



The production of bespoke synthetic teaching datasets without access to the original data

DOI:

[10.1007/978-3-031-69651-0_10](https://doi.org/10.1007/978-3-031-69651-0_10)

Document Version

Accepted author manuscript

[Link to publication record in Manchester Research Explorer](#)

Citation for published version (APA):

Elliot, M., Little, C., & Allmendinger, R. (2024). The production of bespoke synthetic teaching datasets without access to the original data. In *Privacy in Statistical Databases: International Conference, PSD 2024, Antibes Juan-les-Pins, France, September 25–27, 2024, Proceedings* (pp. 144–157). (Lecture Notes in Computer Science; Vol. 14915). Springer Nature. https://doi.org/10.1007/978-3-031-69651-0_10

Published in:

Privacy in Statistical Databases

Citing this paper

Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

General rights

Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy

If you believe that this document breaches copyright please refer to the University of Manchester's Takedown Procedures [<http://man.ac.uk/04Y6Bo>] or contact openresearch@manchester.ac.uk providing relevant details, so we can investigate your claim.



The production of bespoke synthetic teaching datasets without access to the original data

Mark Elliot¹[0000–0002–3142–4493], Claire Little¹[0000–0003–4803–3007], and
Richard Allmendinger²[0000–0003–1236–3143]

¹ School of Social Sciences, University of Manchester, Manchester, M13 9PL, UK
`{mark.elliott,claire.little}@manchester.ac.uk`

² Alliance Manchester Business School, University of Manchester, Manchester, M13
9PL, UK
`richard.allmendinger@manchester.ac.uk`

Abstract. Teaching datasets are a pivotal component of the data discovery pipeline. These datasets often serve as the initial point of interaction for data users, allowing them to explore the contents of a dataset and assess its relevance to their needs. However, there are instances where their viability is limited, particularly where source data is only accessible within restricted settings, such as trusted research environments (TREs). In response to this challenge, this paper proposes the production of synthetic datasets tailored for specific teaching purposes by utilising already cleared (and published) analyses as the basis for the synthesis. Unlike generic synthetic datasets, the datasets created are designed to solely reproduce the specific analyses. Crucially, the datasets can be generated without access to the original data. Two experiments with census data demonstrate the viability of the method and a live use case is described. Issues arising such as marginal disclosure risk are then discussed.

Keywords: Data Synthesis · Evolutionary Algorithms · Data Utility · Disclosure Risk.

1 Introduction

Data Synthesis (DS) [13,17] is a methodology within statistics and machine learning that produces an artificial dataset, that does not contain any real records but approximates the underlying data structure of the original (real) data whilst (hopefully) having low disclosure risk. DS can be used as a data confidentiality method to prevent leakage of confidential information about data subjects whilst delivering analytical utility equivalent to that of the original data. Data synthesis can therefore allow better access to information that might otherwise be safeguarded or restricted since it presents a lower disclosure risk than the original data. For a broad review of DS see [4].

Teaching datasets are a pivotal component of the data discovery pipeline. These datasets are often the initial exposure of users to the data, allowing them

to explore a dataset’s contents and assess its relevance to their needs. Typically constituting compact subsets of the complete dataset, traditional teaching datasets employ data minimisation techniques to control disclosure risks.

Despite this crucial role in the data discovery process, there are instances where the viability of orthodox teaching datasets is limited, particularly in scenarios where the source data is only accessible within restricted settings, such as trusted research environments (TREs). Some TREs have attempted to address this limitation by creating generic synthetic datasets for teaching purposes. However, this approach encounters a significant challenge inherent in all general synthetic data – it may fail to accurately replicate the analytical results desired by trainers for their students. TREs can also be uncomfortable with assessing disclosure risk for synthetic data as it does not map onto standard output disclosure control rules (see e.g. [8])

In response to this challenge, this paper proposes an alternative: the production of bespoke synthetic datasets tailored for specific teaching purposes. The envisioned approach utilises already cleared and published analyses as the basis for the synthesis. This could be the trainer’s own analyses or those published by a third party. Unlike generic synthetic datasets, the bespoke synthetic datasets are designed to solely reproduce the required analyses. The ultimate vision is to enable trainers to produce their own bespoke synthetic teaching datasets that look and feel like real datasets and can reproduce the required outputs faithfully. Crucially, the datasets can be generated without access to the original data.

Access to such bespoke synthetic datasets offers an opportunity for users to undergo training using realistic data before applying for access to the real restricted dataset. This may not only enhance the training experience, but also reduce the time required within a TRE. By introducing this innovative approach to DS, this work aims to redefine the landscape of data accessibility, offering enhanced training experiences and expanded opportunities for data exploration within the research data community.

2 Background

The primary objective of this paper is to investigate the feasibility of generating teaching datasets tailored to restricted data access scenarios. Building on earlier proof-of-concept work [5,12,10], we implement methods from population-based search, and in particular Evolutionary Algorithms (EAs) [16] to achieve this.

DS is a new application for the EA community. Initial proof-of-concept work [5,10,12] using census data has provided a formal definition of the problem (i.e., decision variables, constraints, and objective functions) and then investigated the suitability and robustness of off-the-shelf methods to tackle the problem. A key difference between our proposed EA method and existing data synthesis methods (both the statistical and machine learning variants) is that no access to the original dataset is required to create the synthetic data, making it suitable for our use case. We are not aware of other work that does this; however, we acknowledge that as early as 2003, Burrige [1] proposed a method

called Information Preserving Statistical Obfuscation (IPSO), which generates a dataset that reproduces the original analysis – but this was performed with access to the original data, unlike our approach.

EAs perform iterative optimisation using three main, biologically inspired operators: selection, crossover, and mutation. Briefly: an initial population of candidate solutions is specified. In our case, a candidate solution is a synthetic dataset (using synthetic datasets as candidates differs from existing EA methods that tend to use strings or vectors of data of the same type), and the fitness (quality) of the candidates is calculated using the objective function.

The parental selection operator is used to select candidates (parents) to reproduce for a new population, with fitter candidates more likely to be selected. The crossover operator combines some of the parents to produce new candidate solutions (children). The mutation operator then mutates some of the candidates (i.e., randomly changes some of the decision variables). 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 fitter solutions produced with each generation. Typically, the process terminates when a specified number of generations has been produced or a particular fitness level has been reached.

A useful feature of EAs is their flexibility - there are many parameters that can be changed or set, and the objective function can be designed for the specific purpose (one can also optimize multiple competing objective functions, also known as multi-objective optimization). In this study, the fitness function is designed to optimise the synthetic data such that it matches the desired analytical output. Previous work [2,3] has shown the feasibility of using EAs to generate synthetic microdata.

3 Research Design

3.1 General Approach

The study reported here consists of two experiments (with a live use case described in Section 6).

In the case of two experiments, an output or set of outputs from a dataset are defined. The dataset is split into two (original and holdout) and the analyses are then run on the original dataset to produce the *reference outputs*.

Those reference outputs are then converted into objective functions for the EA, which produces the synthetic dataset. The EA is then run until it converges. The analytical output is then produced for the synthetic data and compared to the reference output and the equivalent output of the analyses applied to the holdout data. The holdout data simulates another sample from the same population as the original.

Since the method does not access the original dataset, it requires analytical output from the original data that has already been cleared as safe for publication. This could be coefficients of a regression model or summary statistics, for

example. It also requires basic metadata; the names of each of the variables and the possible values they can take.

Data The data used were a subset of the 1991 UK Census Sample of Anonymised Records [15]. Analytic output is based on a paper by Gardiner and Hill [7] – which examined car ownership of the elderly (aged ≥ 50) population of Sheffield. We have expanded this (to allow for a larger dataset) and include all of the UK county of South Yorkshire (which includes Sheffield). The base dataset consists of 8054 individuals aged 50 or over in the South Yorkshire area who had answered the car question (how many cars, if any, they had access to). This base dataset was split into an experimental dataset (henceforth referred to as the “original dataset”)³ and a holdout dataset both containing 4027 records, with the holdout dataset not used at all in the creation of the synthetic data. The variables used were: AREAP (a geocode), AGE (in single years), SEX, ETHGROUP (ethnic group), LTILL (the presence of long-term illness), TENURE, CARS (number of cars that the respondent has access to). The univariate distributions of these variables are shown in Table 1.

Table 1: Univariate distributions of the variables used in the study derived from the original dataset.

Name	Value	Label	Proportion
AREAP	Barnsley	48	0.1716
	Doncaster	49	0.2210
	Rotherham	50	0.1838
	Sheffield	51	0.4236
AGE	50-95		
SEX	Male	1	0.4527
	Female	2	0.5473
ETHGROUP	White	1	0.9878
	Other	2	0.0122
LTILL	Yes	1	0.3700
	No	2	0.6300
TENURE	Own occ. outright	1	0.3650
	Own occ. Buying	2	0.2131
	Rented priv. furn.	3	0.0055
	Rented job/business	5	0.0117
	Rented housing assoc.	6	0.0204
	Rented local authority	7	0.3486
CARS	No access	0	0.4726
	Access to ≥ 1	1	0.5274

³ we use the term original here rather than the more orthodox “training” to make it clear that we are not training a model on the data

4 Experiment 1: A single output

For experiment 1 we focus on just one analytical output - a logistic regression model. To generate the reference output, the logistic regression was performed on the original dataset, using a binary version of CARS as the target variable (where 0=no access to cars and 1=access to 1 or more cars). The reference level for the explanatory variables are: from the Sheffield (AREAP=51) area, white (ETHGROUP=1), male (SEX=1), no long-term illness (LTILL=2) and living in a home that is owned outright (TENURE=1). The Model coefficients for the reference output are shown in Table 2.

Table 2: Logistic regression model coefficients generated from the original data in Experiment 1, where CARS (access to cars) is the target variable.

Parameter	Coefficient
Constant	6.4034
AGE	-0.0787
AREAP_48	-0.0239
AREAP_49	-0.0398
AREAP_50	-0.0650
SEX_2	-0.5501
ETHGROUP_2	-0.8254
LTILL_1	-0.2033
TENURE_2	0.1788
TENURE_3	-1.6399
TENURE_4	-1.7013
TENURE_5	-0.7950
TENURE_6	-1.8219
TENURE_7	-1.9224

4.1 Method

The EA used the coefficients from the logistic regression as its objectives and calculated the mean squared error (MSE) between the original coefficients and coefficients generated using the synthetic data to form the fitness measure.⁴ The EA was run for 2000 generations, with no crossover and a gradually decreasing

⁴ For these experiments we used simple unweighted MSEs (within each output). We acknowledge that other alternatives could have been used and in a separate of experiments we have used standardised coefficients instead – the results were broadly similar. The optimum weighting methodology will be decided by intensive benchmarking work - the goal here was to examine what could be achieved with a very simple algorithm.

mutation rate (this has been found to be optimal in other experiments), starting at 0.01 and decreasing every 250 generations (to 0.005, 0.001, 0.0005, 0.0001, . . .). The population size was 24 (i.e. after the selection process for each generation there are 24 candidate synthetic datasets) and the initial population was generated using the uniform distribution. When a value is mutated, its replacement is drawn from the univariate distribution (from Table 1). An elitist selection strategy was used in which the best children/parents of each generation form the next generation. Five randomly initialised (with a different random seed each time) runs were performed.

4.2 Results

Table 3 shows the regression coefficients for the model produced by the synthetic data in each of the five runs. As can be seen in the table, in each run of the EA the data have converged on a solution very close to the original dataset.

In Table 4 we compare the results obtained by the EA synthesised dataset (from run 1) with those of the holdout dataset. The coefficients for the EA synthesised dataset are much closer to those of the original dataset than those of the holdout dataset are. To be clear about what this implies: *this method produces more accurate reproduction of the required analytical properties than a second sample drawn from the same population (using the same sampling mechanism)*. For reference, Table 4 also shows the results obtained by a general synthetic dataset produced from the original (using CART in synthpop [14] with the default settings).

5 Experiment 2: Multiple outputs

In the second experiment, we added two additional outputs: cross-tabulations between the CARS variable and two other variables (TENURE and ETHGROUP); these can be found in Tables 5 and 6. This mimics exploratory analysis that might be conducted as part of the training exercise before the model itself is run.

5.1 Method

As in Experiment 1, the EA was run for 2000 generations, without crossover and a gradually decreasing mutation rate, starting at 0.01 and decreasing every 250 generations (to 0.005, 0.001, 0.0005, 0.0001, . . .). There were 24 synthetic datasets in the population, and the initial population was generated using the uniform distribution. When mutating a value, its replacement is drawn from the univariate distribution (from Table 1). An elitist strategy was used, in which the optimal children/parents of each generation form the next.

The EA used the three outputs as objectives and calculated the mean squared error (MSE) between those and the outputs generated using the synthetic data. A weighted sum of the three outputs drove the EA. Ten different weightings were tried: all equal and then slowly decreasing the weighting given to the logistic

Table 3: A comparison of model coefficients for the original and synthetic data for five runs of the synthesiser.

Parameter	Original output coefficient	Synthetic data coefficient				
		Run1	Run 2	Run 3	Run 4	Run 5
Constant	6.4034	6.4034	6.4032	6.4035	6.4036	6.4037
AGE	-0.0787	-0.0787	-0.0752	-0.0745	-0.0731	-0.0743
AREAP_48	-0.0239	-0.0239	-0.0239	-0.0241	-0.0238	-0.0239
AREAP_49	-0.0398	-0.0398	-0.0397	-0.0398	-0.0400	-0.0398
AREAP_50	-0.0650	-0.0650	-0.0649	-0.0646	-0.0646	-0.0650
SEX_2	-0.5501	-0.5501	-0.5500	-0.5498	-0.5500	-0.5504
ETHGROUP_2	-0.8254	-0.8254	-0.8253	-0.8252	-0.8256	-0.8254
LTILL_1	-0.2033	-0.2033	-0.2032	-0.2033	-0.2034	-0.2035
TENURE_2	0.1788	0.1788	0.1789	0.1786	0.1787	0.1789
TENURE_3	-1.6399	-1.6399	-1.6404	-1.6394	-1.6394	-1.6407
TENURE_4	-1.7013	-1.7031	-1.7033	-1.7033	-1.7030	-1.7025
TENURE_5	-0.7950	-0.7950	-0.7949	-0.7948	-0.7950	-0.7949
TENURE_6	-1.8219	-1.8219	-1.8217	-1.8219	-1.8217	-1.8217
TENURE_7	-1.9224	-1.9224	-1.9225	-1.9226	-1.9227	-1.9220
Mean absolute error between		0.00037	0.00049	0.00059	0.00053	0.00053
run and original output:						

Table 4: A comparison of model coefficients for the original, EA synthetic, a general synthetic dataset and the holdout dataset.

	Regression coefficients			
	Original	EA synthetic	General synthetic	Holdout
Constant	6.4034	6.4034	6.6741	5.7832
AGE	-0.0787	-0.0787	-0.0813	-0.0707
AREAP_48	-0.0239	-0.0239	-0.1057	-0.0345
AREAP_49	-0.0398	-0.0398	-0.1767	-0.1533
AREAP_50	-0.0650	-0.0650	-0.2197	0.1968
SEX_2	-0.5501	-0.5501	-0.5176	-0.6311
ETHGROUP_2	-0.8254	-0.8254	0.8265	-0.8528
LTILL_1	-0.2033	-0.2033	-0.2442	-0.0914
TENURE_2	0.1788	0.1788	-0.0391	0.3443
TENURE_3	-1.6399	-1.6399	-1.7678	-1.8901
TENURE_4	-1.7013	-1.7031	-1.6062	-1.5852
TENURE_5	-0.7950	-0.7950	-0.6090	-0.5415
TENURE_6	-1.8219	-1.8219	-2.0131	-1.8200
TENURE_7	-1.9224	-1.9224	-1.9847	-1.8996
Mean absolute error between		0.0004	0.2323	0.1460
dataset and original:				

Table 5: Cross-tabulation of tenure and access to car derived from original dataset.

TENURE	Frequency		Proportion	
	No access to car	Access to car	No access to car	Access to car
Own occ. outright	470	1000	0.117	0.248
Own occ. Buying	162	696	0.04	0.173
Rented priv. furn.	15	7	0.004	0.002
Rented priv. unfurn.	105	39	0.026	0.01
Rented job/business	22	25	0.005	0.006
Rented housing assoc.	64	18	0.016	0.004
Rented local authority	1065	339	0.264	0.084

Table 6: Cross-tabulation of ethnic group and access to car derived from original dataset.

ETHGROUP	Frequency		Proportion	
	No access to car	Access to car	No access to car	Access to car
White	1880	2098	0.467	0.521
Other	23	26	0.006	0.006

regression output and increasing the weighting for the two table outputs (they are labelled as run 1-10), the weightings are listed in Appendix A. ⁵.

5.2 Results

Unlike the first experiment, the runs of the EA differ by varying by the weights of the different objectives. In Table 7 we compare the results obtained by the EA synthetic dataset for runs, 1, 3, 5, 7 and 10 with each other, the holdout dataset and a general synthetic dataset. As can be seen, regardless of the setting of the weight parameter, the EA synthetic data are closer to the original data than the holdout dataset in reproducing the model. Although they are further from the original than in experiment 1 - particularly with run 1 - the differences are still markedly less than those that would be produced by a second equivalent sample and would therefore we assume, be acceptable (for the use case).

Equivalent results for the cross-tabulations can be found in Table 8 and 9. Here run 10 outperforms the others - unsurprisingly, given it was weighted to optimising these outputs.

⁵ Since the MSE of the regression results was higher than those for the table outputs, weighting it lower had the effect of avoiding the regression dominating the outcome

Table 7: Comparison of logistic model output of the original data with runs 1, 3, 5, 7 and 10 of the EA synthetic data and the holdout data.

	Original	EA Run 1	EA Run 3	EA Run 5	EA Run 7	EA Run 10	Holdout	General Synth.
Constant	6.4034	6.4016	6.4074	6.3980	6.3944	6.3215	5.7832	6.6741
AGE	-0.0787	-0.0729	-0.0745	-0.0754	-0.0760	-0.0773	-0.0707	-0.0813
AREAP_48	-0.0239	-0.0201	-0.0236	-0.0293	-0.0222	-0.0650	-0.0345	-0.1057
AREAP_49	-0.0398	-0.0379	-0.0371	-0.0388	-0.0339	-0.0779	-0.1533	-0.1767
AREAP_50	-0.065	-0.0650	-0.0631	-0.0662	-0.0655	-0.0922	0.1968	-0.2197
SEX_2	-0.5501	-0.5539	-0.5496	-0.5491	-0.5477	-0.3885	-0.6311	-0.5176
ETHGROUP_2	-0.8254	-0.8201	-0.8114	-0.8223	-0.8228	-0.8061	-0.8528	0.8265
LTILL_1	-0.2033	-0.2043	-0.2019	-0.2052	-0.2086	-0.2708	-0.0914	-0.2442
TENURE_2	0.1788	0.1739	0.1790	0.1795	0.2029	0.6984	0.3443	-0.0391
TENURE_3	-1.6399	-1.6366	-1.6330	-1.6359	-1.6308	-1.6154	-1.8901	-1.7678
TENURE_4	-1.7013	-1.7023	-1.6993	-1.7092	-1.7009	-1.7676	-1.5852	-1.6062
TENURE_5	-0.795	-0.7934	-0.7914	-0.7876	-0.7984	-0.8221	-0.5415	-0.609
TENURE_6	-1.8219	-1.8145	-1.8208	-1.8200	-1.8252	-1.9509	-1.82	-2.0131
TENURE_7	-1.9224	-1.9276	-1.9308	-1.9452	-1.9704	-2.2264	-1.8996	-1.9847
Mean absolute error: between dataset & original:		0.0034	0.0037	0.0048	0.0085	0.1078	0.1460	0.2323

Table 8: A comparison of the proportion of respondents with access to a car conditioned on tenure in the original data, runs 1, 3, 5, 7 and 10 of the EA synthetic data, general synthetic and the holdout data.

TENURE	Original	EA Run 1	EA Run 3	EA Run 5	EA Run 7	EA Run 10	Holdout	General Synth.
Own occ. outright	0.6795	0.6985	0.6923	0.6805	0.6785	0.6796	0.6582	0.6929
Own occ. Buying	0.8122	0.7436	0.7291	0.7300	0.7164	0.8028	0.8145	0.7952
Rent priv. furn.	0.3333	0.4444	0.3871	0.3846	0.3846	0.3750	0.2500	0.2500
Rent priv. unfurn.	0.2778	0.4035	0.3800	0.3462	0.3617	0.3056	0.2703	0.2903
Rent job/business	0.5455	0.5758	0.5313	0.5517	0.5357	0.5385	0.5625	0.5455
Rent housing assoc.	0.2000	0.3571	0.3830	0.3333	0.3429	0.2857	0.2273	0.1739
Rent local authority	0.2414	0.3458	0.3205	0.3079	0.2914	0.2399	0.2281	0.2443
Mean absolute error: between dataset & original:		0.0880	0.0755	0.0584	0.0621	0.0247	0.0246	0.0222

6 Discussion

The results described above are compelling. They show how it is possible to produce usable synthetic data that is able to replicate multiple analytical outputs at a level better than that produced by another sample from the same population.

Table 9: A comparison of the proportion of respondents with access to car conditioned on tenure in the original data, to runs 1, 3, 5, 7 and 10 of the EA synthetic data, general synthetic and the holdout data.

ETHGROUP	Original	EA Run 1	EA Run 3	EA Run 5	EA Run 7	EA Run 10	Holdout	General Synth.
White	0.5273	0.5613	0.5480	0.5359	0.5273	0.5273	0.5162	0.5254
Other	0.5000	0.4430	0.4247	0.4394	0.4286	0.4615	0.5385	0.7143
Mean absolute error:		0.0455	0.0480	0.0346	0.0357	0.0192	0.0248	0.1081

This then provides an in principle mechanism for trainers to produce their own teaching datasets with access to the original data.

We were also interested in testing how easy it would be to produce an actual teaching dataset of sufficient quality to be used by a trainer in practice. One of the requirements this threw up was the addition of other statistical properties beyond a single model (further descriptive statistics and weights). Working with Administrative Data Research UK we developed a teaching dataset of the UK linked ASHE-Census dataset⁶. The dataset produced was generated on the basis of outputs in the methodology report produced by the team that developed the ASHE-Census dataset [6]. The requirements were that the data should be able to reproduce a linear regression model predicting income, the univariate frequencies of five variables (sex, ethnicity, education level and disability, UK born) and the conditional means and medians of the income variable on those five variables. One additional step was required; the addition of weights. For the purposes of providing a realistic training experience, a weight variable needed to be added to mimic what the trainees would find when they eventually had access to the real data in the TRE. The weight was created by calibrating the synthetic dataset to census microdata.

The synthetic dataset was successfully produced to the satisfaction of the trainer and was first used in a training course run by the UK’s National Centre for Research Methods in April 2024⁷.

6.1 Disclosure risk

One question that might arise when considering this method is what about the disclosure risk? Returning to experiment 1, we carried out some analyses which were helpful indicators.

Firstly, the most obvious place for an adversary to attack this data would be to use the explanatory variables in an attempt to disclose the values of the

⁶ See: <https://www.adruk.org/news-publications/news-blogs/new-linked-dataset-available-to-provide-insights-into-earnings-and-employment-in-england-and-wales/>

⁷ <https://www.ncrm.ac.uk/training/show.php?article=13306>

response variable in the model used to construct the data. We used the TCAP statistic [9,10,11] to assess this. See Table 10.

Table 10: TCAP values for the various datasets used in experiment 1 given the target is the response variable CARS and the keys are the explanatory variables.

Dataset comparison	Raw TCAP	Calibrated TCAP
Original -> Original	1.000	1.000
Original -> EA run 1	0.733	0.464
Holdout -> EA run 1	0.719	0.437
Original -> General Synthetic	0.777	0.552
CAP Baseline	0.502	0.000

TCAP assesses the probability of an adversary that links a known population unit to some data correctly inferring the target attribute for that population given that the l -diversity = 1, given the values of key variables. The baseline value is that provided when the adversary makes a random draw from the univariate distribution of the target (essentially guessing). Unsurprisingly, the synthetic data provides a better inferability than guessing. However, there are two things to note here. The adversary is almost as likely to be able to make correct inferences about population units that are **not** in the data that have been used to produce the analytical output (as represented by the holdout data) than about those that are represented in that data. The TCAP value here therefore represents the inference risk arising from the model itself. This is intuitive. Since we have used the model and no other information to generate the synthetic data, it makes sense that any risk identified is the risk arising from the model itself. However, we can dig a little deeper here. It is possible that the nature of an evolutionary algorithm means that secondary analytical properties may emerge from the process. We can examine this by using TCAP to treat every other variable in the dataset as a possible target. This analysis is shown in Table 11.

The key take away from this is that the synthetic data is barely more informative than the baseline and in this case is slightly more informative about the holdout data than the original data. Essentially, there is no emergent risk in the data beyond that which is already present in the model itself. Since in our scenario here the model will have already cleared and published, we can say that in this case the synthetic data would present negligible marginal risk. Note that this is not a general conclusion about this mechanism - data would have to be assessed on a case by case basis. Particularly where there are multiple outputs, it is possible to imagine cases where unanticipated marginal risk might emerge.

Table 11: TCAP values for the various datasets used in experiment 1 given the target is each of the explanatory variables and the keys are the remaining variables in the dataset.

	Target variable:						Mean
	AREAP	AGE	SEX	ETHGROUP	LTILL	TENURE	
Original -> Original	1	1	1	1	1	1	1
Original -> EA run 1	0.296	0	0.543	0.953	0.583	0.367	0.457
Holdout -> EA run 1	0.299	0	0.516	0.949	0.576	0.377	0.453
Original -> General Synth.	0.481	0.194	0.682	0.984	0.744	0.569	0.609
Baseline	0.292	0.029	0.504	0.976	0.534	0.302	0.440

6.2 Concluding remarks

The paper has presented a new approach to generating synthetic data based on analytical output without requiring access to real data. The particular use case here is the production of teaching dataset, and for this it appears very promising.

In terms of the experiments reported here, the engaged reader will have noted that the outputs did not include standard errors, fit measures, deviance scores and so on that one would typically expect in standard statistical output. Understanding the impact of multiple outputs on the data quality and how to manage complexity of that with adaptive mutation, diversity management and other more advanced algorithms is the focus of our current work.

However, the method we have demonstrated also opens up another, perhaps more intriguing, possibility. By embedding an analytical output in a synthetic dataset the method opens up the possibility of formalising the assessment of disclosure risk for analytical outputs from safe settings. At present, this process is typically managed through the combination of the application of rules and the evaluation of a human output checker, and so moving this on to more formal grounds (perhaps in the form of a toolkit for output checker to use) would be an advance. But the potential is that by embedding the output in synthetic data it is possible to assess the risk using standard microdata disclosure control methods as we have done here. This is further element of our current work.

A second possible extension of this work is the synthesising of data from unlinked sources. It is a modest extension of the experiment 2 above to imagine multiple datasets (perhaps drawn from the same population) from which analytical outputs have been produced being quasi-linked through the production of a joint synthetic dataset without anyone even having sight of both datasets.

In future work, we will be exploring both of these possibilities as well as considering the application of the method reported here to the production of a wider range of teaching datasets.

Acknowledgments. This study was funded by the Economic and Social Research Council (grant numbers ES/Z502984/1 and ES/T000066/1). We also thank Lucy Stokes

of the National Institute of Economic and Social Research for her input into the ASHE-Census live test.

Code. The code for the experiments reported here can be found at: <https://github.com/claireliddle/psd2024-bespoke-synthetic-datasets>

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Burrige, J.: Information preserving statistical obfuscation. *Statistics and Computing* **13**, 321–327 (2003). <https://doi.org/10.1023/A:1025658621216>
2. Chen, Y., Elliot, M., Sakshaug, J.: Genetic algorithms in matrix representation and its application in synthetic data. In: UNECE Worksession on Statistical Confidentiality 2017 (2017), https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2017/2_Genetic_algorithms.pdf
3. Chen, Y., Elliot, M., Smith, D.: The application of genetic algorithms to data synthesis: A comparison of three crossover methods. In: Domingo-Ferrer, J., Montes, F. (eds.) *Privacy in Statistical Databases*. pp. 160–171. Springer International Publishing (2018). https://doi.org/10.1007/978-3-319-99771-1_11
4. Drechsler, J., Haensch, A.C.: 30 years of synthetic data (2023). <https://doi.org/10.48550/arXiv.2304.02107>
5. Elliot, M., Little, C., Allmendinger, R.: Do samples taken from a synthetic microdata population replicate the relationship between samples taken from an original population? In: UNECE Expert Meeting on Statistical Data Confidentiality 2023 (2023), https://unece.org/sites/default/files/2023-08/SDC2023_S4_5_UnivManchester_Elliot_D.pdf
6. Forth, J., Phan, V., Ritchie, F., Whittard, D., Stokes, L., Bryson, A., Singleton, C.: ASHE – census 2011 data linkage: User Guide for Drop 2 of the ASHE-Census 2011 Dataset (2022), <https://www.wagedynamics.com/wp-content/uploads/2023/01/ASHE-CEW11-User-Guide-Version-2.1-Drop-2.pdf>
7. Gardiner, C., Hill, R.: Analysis of access to cars from the 1991 uk census samples of anonymised records: a case study of the elderly population of sheffield. *Urban Studies* **33**(2), 269–281 (1996)
8. Griffiths, E., C., G., Kotrotsios, Y., Parker, S., Scott, J., Welpton, R., Wolters, A., Woods, C.: *Handbook on Statistical Disclosure Control for Outputs* (2019), https://ukdataservice.ac.uk/app/uploads/thf_datareport_aw_web.pdf
9. Jennifer, T., Mark, E.: The synthetic data challenge. Conference of European Statisticians (2019), https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2019/mtg1/SDC2019_S3_UK_Synthetic_Data_Challenge_Elliot_AD.pdf
10. Little, C., Elliot, M., Allmendinger, R.: Comparing the utility and disclosure risk of synthetic data with samples of microdata. In: *Privacy in Statistical Databases*. pp. 234–249. Springer International Publishing, Paris, France, September 21–23, 2022 (2022). https://doi.org/10.1007/978-3-031-13945-1_17
11. Little, C., Elliot, M., Allmendinger, R.: Synthetic census microdata generation: A comparative study of synthesis methods examining the trade-off between disclosure risk and utility. *Journal of Official Statistics* (2024)

12. Little, C., Elliot, M., Allmendinger, R., Samani, S.S.: Generative Adversarial Networks for Synthetic Data Generation: A Comparative Study. In: Joint UN-ECE/Eurostat Expert Meeting on Statistical Data Confidentiality (2021), https://unece.org/sites/default/files/2021-12/SDC2021_Day2_Little_AD.pdf
13. Little, R.J.A.: Statistical Analysis of Masked Data. *Journal of Official Statistics* **9**(2), 407–426 (1993)
14. Nowok, B., Raab, G., Dibben, C.: Synthpop: Bespoke creation of synthetic data in R. *Journal of Statistical Software* **74**(11) (2016). <https://doi.org/10.18637/jss.v074.i11>
15. Office for National Statistics, Census Division, University of Manchester, Cathie Marsh Centre for Census and Survey Research: Census 1991: Individual Sample of Anonymised Records for Great Britain (SARs) (2013). <https://doi.org/10.5255/UKDA-SN-7210-1>
16. Reeves, C., Rowe, J.E.: Genetic algorithms: principles and perspectives: a guide to GA theory, vol. 20. Springer Science & Business Media (2002)
17. Rubin, D.B.: Statistical Disclosure Limitation. *Journal of Official Statistics* **9**(2), 461–468 (1993)

Appendices

Appendix A

Table 12: Objective function weights for the outputs used in Experiment 2.

Run	Output 1 weight	Output 2 weight	Output 3 weight
1	0.33	0.33	0.33
2	0.25	0.375	0.375
3	0.2	0.4	0.4
4	0.15	0.425	0.425
5	0.1	0.45	0.45
6	0.05	0.475	0.475
7	0.025	0.4875	0.4875
8	0.01	0.495	0.495
9	0.001	0.4995	0.4995
10	0.0001	0.49995	0.49995