On integrating the number of synthetic data sets m into the *a priori* synthesis approach

James Jackson[1][0000−0002−4832−6638], Robin Mitra[2][0000−0001−9584−8044], Brian Francis[1][0000−0001−7926−9085], and Iain Dove[3][0000−0002−1145−2999]

[1] Lancaster University, Lancaster, UK
[2] Cardiff University, Cardiff, UK
[3] Office for National Statistics, Titchfield, UK

Abstract. The synthesis mechanism given in [4] uses saturated models, along with overdispersed count distributions, to generate synthetic categorical data. The mechanism is controlled by tuning parameters, which can be tuned according to a specific risk or utility metric. Thus expected properties of synthetic data sets can be determined analytically *a priori*, that is, before they are generated. While [4] considered the case of generating $m = 1$ data set, this paper considers generating $m > 1$ data sets. In effect, m becomes a tuning parameter and the role of m in relation to the risk-utility trade-off can be shown analytically. The paper introduces a pair of risk metrics, $\tau_3(k, d)$ and $\tau_4(k, d)$, that are suited to $m > 1$ data sets; and also considers the more general issue of how best to analyse $m > 1$ categorical data sets: average the data sets pre-analysis or average results post-analysis. Finally, the methods are demonstrated empirically with the synthesis of a constructed data set which is used to represent the English School Census.

Keywords: synthetic data · privacy · categorical data · risk metrics · contingency tables

1 Introduction

When disseminating data relating to individuals, there are always two conflicting targets: maximising utility and minimising disclosure risk. To minimise risk, statistical disclosure control (SDC) methods, which typically involve either suppressing or perturbing certain values, are applied to a data set prior to its release. One such method is the generation of synthetic data sets [14, 6], which involves simulating from a model fit to the original data. These methods, while reducing risk, adversely impact the data's utility resulting in a clear trade-off between risk and utility.

This paper focuses on the role of multiple data sets when synthesizing categorical data (that is, data consisting of only categorical variables) at the aggregated level using saturated count models [4]. Saturated synthesis models allow the synthesizer to generate synthetic data with certain pre-specified properties, thus allowing them to easily tailor the synthesis to suit the data environment

[3]. For example, if the intention is to release open data, relatively more noise can be applied to the data than if the data are released in a secure environment. While the Poisson model is often used to model categorical data, for synthesis this is not necessarily an optimal choice, because the synthesizer - that is, the person(s) responsible for synthesizing the data - has no control over the variance and has, therefore, no way to add additional noise to at-risk records in the data. For this reason, the negative binomial (NBI), a two-parameter count distribution, is much more effective for synthesis. As the NBI distribution's variance is not completely determined by the mean - though the variance is always greater than the mean - the variance *can* be increased accordingly. Nevertheless, there are still restrictions and these are discussed later on.

Specifically, this paper explores how flexibility can be incorporated into the mechanism through the use of multiple synthetic data sets. In some cases (as explained in Section 3), $m > 1$ synthetic data sets must be generated; while in other cases, though it may be sufficient to generate just $m = 1$ synthetic data set, the optimal m can still be considered in relation to the risk-utility trade-off: does the improvement in utility sufficiently outweigh the cost in terms of greater risk? This is because, since it reduces simulation error, increasing m leads to greater utility but also, inevitably, greater risk [11, 12]. More generally, considering $m > 1$ introduces another tuning parameter for the synthesizer to set, thereby providing further flexibility.

This paper is structured as follows: Section 2 summaries the (σ, α)-synthesis mechanism, on which the results in this paper are based; Section 3 extends the mechanism to incorporating $m > 1$; Section 4 introduces the $\tau_3(k, d)$ and $\tau_4(k, d)$ metrics, developed to assess risk in multiple categorical synthetic data sets; Section 5 presents an illustrative example; and lastly Section 6 ends the paper with a discussion and areas of future research.

2 Review of the use of saturated models for synthesis

The discrete nature of categorical data allow it to be expressed as a multi-dimensional contingency table (multi-way table). As a multi-way table, the data consist of a structured set of cell counts $f_1, \ldots, f_K$, which give the frequencies with which each combination of categories is observed.

Synthetic data sets can then be generated by replacing these observed counts (known henceforth as "original counts") with synthetic counts. There are two distinct modelling methods for contingency tables: multinomial models and count models. The multinomial approach ensures that the total number of individuals in the original data n is equal to the total number of individuals in the synthetic data n_{syn}. The `syn.catall` function in the R package **synthpop** [7] can be used to generate synthetic data via a saturated multinomial model.

The (σ, α)-synthesis mechanism [4] uses saturated *count* models for synthesis; specifically, either a saturated negative binomial (NBI) model or a saturated Poisson-inverse Gaussian (PIG) [13] model. In this paper, for brevity, only the

NBI has been considered. Besides, the NBI and PIG distributions are broadly similar, as they share the same mean-variance relationship.

The (σ, α)-synthesis mechanism has two parameters which are set by the synthesizer. The first, $\sigma > 0$, is the scale parameter from a two-parameter count distribution (such as the NBI). The parameter σ can be tuned by the synthesizer to adjust the variability in the synthetic counts, thus increasing or decreasing their expected divergence from the original counts. More noise is required for sensitive cells - usually small cell counts, which correspond to individuals who have a unique (or near-unique) set of observations - to generate sufficient uncertainty to mask the original counts' true values.

The mechanism's second parameter, denoted by $\alpha \geq 0$, relates to the size of the pseudocount - in practice, this is not actually a count but a small positive number such as 0.01 - which is added to zero cell counts (zero cells) in the original data. This assigns a non-zero probability that a zero cell is synthesized to a non-zero. The pseudocount α is only applied to non-structural zero cells (known as random or sampling zeros), which are zero cells for which a non-zero count *could* have been observed. Throughout this paper it has been assumed, for brevity, that $\alpha = 0$.

Given an original count $f_i = N_i$ $i = 1, \ldots, K$, the corresponding synthetic count f_i^{syn} is drawn from the following model:

$$f_i^{\text{syn}} \mid f_i = N_i, \sigma \sim \text{NBI}(N_i, \sigma), \quad \text{and therefore,}$$

$$p(f_i^{\text{syn}} = N_2 \mid f_i = N_1, \sigma) = \frac{\Gamma(N_2 + 1/\sigma)}{\Gamma(N_2 + 1) \cdot \Gamma(1/\sigma)} \cdot \left(\frac{\sigma N_1}{1 + \sigma N_1}\right)^{N_2} \cdot \left(\frac{1}{1 + \sigma N_1}\right)^{1/\sigma}.$$

Using a saturated count model has certain advantages in data synthesis. Firstly, it guarantees the preservation of relationships between variables, as no assumptions are made as to which interactions exist. Secondly, the method scales equally well to large data sets, as no model fitting is required - the model's fitted counts are just the observed counts. Finally, as the fitted counts are just equal to the observed counts, it allows expected properties of the synthetic data to be determined *a priori* (that is, prior to synthesis). The (unwelcome) uncertainty around model choice is, in effect, minimised, and instead uncertainty is injected where it is most needed: to add noise to sensitive cells in the original data.

2.1 The τ metrics

The following τ metrics [4], give a basic quantification of risk (and utility) in tabular data:

$$\tau_1(k) = p(f^{\text{syn}} = k) \qquad \tau_3(k) = p(f^{\text{syn}} = k | f = k)$$
$$\tau_2(k) = p(f = k) \qquad \tau_4(k) = p(f = k | f^{\text{syn}} = k),$$

where f and f^{syn} are arbitrary original and synthetic counts, respectively. The metric $\tau_2(k)$ is the empirical proportion of *original* counts with a count of k, and $\tau_1(k)$ is the proportion of *synthetic* counts of size k. The metric $\tau_3(k)$ is the

probability that an original count of size k is synthesized to k; and $\tau_4(k)$ is the probability that a synthetic count of size k originated from a count of size k. The metrics $\tau_3(1)$ and $\tau_4(1)$, in particular, are the most associated with risk, as these relate to uniques and can be viewed as outliers in the data. When, for example, $\tau_4(1)$ is close to 1, it is possible to identify, with near certainty, uniques in the original data from the synthetic data.

When saturated models are used, the expected values of these τ metrics can be found analytically as functions of the tuning parameters (σ, α and, as later described, m). Hence the synthesizer knows, *a priori*, the noise required to achieve a given $\tau_3(1)$ or $\tau_4(1)$ value.

3 The role of m as a tuning parameter

The original inferential frameworks for fully and partially synthetic data sets [9, 10] relied on the generation of $m > 1$ synthetic data sets, because they required the computation of the between-synthesis variance b_m (see below). However, when the original data constitute a simple random sample, and the data are completely synthesized, valid inferences can be obtained from $m = 1$ synthetic data set [8]. In this instance, while $m > 1$ data sets are not intrinsic to obtaining *valid* inferences, the *quality* of inferences - for example, the width of confidence intervals - can, nevertheless, be improved upon by increasing m - but at the expense of higher risk. It is less a question, therefore, of which m allows valid inferences to be obtained, but rather a question of which value of m is optimal with respect to the risk-utility trade-off?

Thus m can be viewed as a tuning parameter, and, as with the other tuning parameters σ and α, expected risk and utility profiles can be derived analytically, *a priori*. When saturated models are used for synthesis, ignoring the small bias arising from $\alpha > 0$, simulation error is the only source of uncertainty - and increasing m reduces simulation error. The notion is that $m > 1$ may allow a more favourable position in relation to the risk-utility trade-off than when $m = 1$; in short, it increases the number of options available to the synthesizer.

The use of parallel processing can substantially reduce the central processing unit (CPU) time when generating multiple data sets. Besides, the CPU time taken is typically negligible anyway; the synthesis presented in Section 5 took 0.3 seconds for the NBI with $m = 1$ on a typical laptop running R.

3.1 Obtaining inferences from $m > 1$ data sets

Analysing the $m > 1$ data sets before averaging the results When analysing multiple synthetic data sets, traditionally the analyst considers each data set separately before later combining inferences. While point estimates are simply averaged, the way in which variance estimates are combined depends on the type of synthesis carried out: such as whether fully or partially synthetic data sets are generated and also whether synthetic counts are generated by simulating from the Bayesian posterior predictive distribution or by simulating

directly from the fitted model. The combining rules also depend on whether an analyst is using the synthetic data to estimate a population parameter Q, or an observed data estimate $\hat{Q}$: the former needs to account for the sampling uncertainty in the original data whereas the latter does not.

Suppose, then, that an analyst wishes to estimate a univariate population parameter Q from $m > 1$ synthetic data sets. A point estimate $q^{(l)}$, and its variance estimate $v^{(l)}$, is obtained from each synthetic data set, $l = 1, \ldots, m$. Before these estimates are substituted into a set of combining rules, it is common, as an intermediary step, to first calculate the following three quantities [2]:

$$\bar{q}_m = \frac{1}{m}\sum_{l=1}^{m} q^{(l)}, \qquad b_m = \frac{1}{(m-1)}\sum_{l=1}^{m}(q^{(l)} - \bar{q}_m)^2, \qquad \bar{v}_m = \frac{1}{m}\sum_{l=1}^{m} v^{(l)},$$

where $\bar{q}_m$ is the mean estimate, b_m is the 'between-synthesis variance', that is, the sample variance of the $m > 1$ estimates, and $\bar{v}_m$ is the mean 'within-synthesis variance', the mean of the estimates' variance estimates.

The quantity $\bar{q}_m$ is an unbiased estimator for $\hat{Q}$, and is so regardless of whether fully or partially synthetic data sets are generated. When using the synthesis method described in Section 2, partially - rather than fully - synthetic data sets are generated, because a synthetic population is not constructed and sampled from, as stipulated in [9]. Hence, the following estimator T_p [10], is valid when estimating $\mathrm{Var}(\hat{Q})$,

$$T_p = \frac{b_m}{m} + \bar{v}_m.$$

The sampling distribution (if frequentist) or posterior distribution (if Bayesian) of $\hat{Q}$ is a t-distribution with $\nu_p = (m-1)\left(1 + m\bar{v}_m/b_m\right)^2$ degrees of freedom. Often, ν_p is large enough for the t-distribution to be approximated by a normal distribution. However, when the between-synthesis variability is much larger than the within-synthesis variability, that is, when b_m is much larger than $\bar{v}_m$ - as may happen when large amounts of noise are applied to protect sensitive records - then ν_p is crucial to obtaining valid inferences.

As the data sets are completely synthesized in the sense of [8] - that is, no original values remain - the following estimator T_s is valid, too, under certain conditions:

$$T_s = \bar{v}_m\Big(\frac{n_{\mathrm{syn}}}{n} + \frac{1}{m}\Big) \approx \bar{v}_m\Big(1 + \frac{1}{m}\Big).$$

These conditions are: firstly, that the original data constitute a simple random sample - therefore, T_s would not be valid if the data originate from a complex survey design - and secondly, that the original data are large enough to support a large sample assumption. The overriding advantage of T_s is that, assuming its conditions do indeed hold, it allows valid variance estimates to be obtained from $m = 1$ synthetic data set.

The large sample assumption facilitates the use of a normal distribution for the sampling distribution (or the posterior distribution) of $\hat{Q}$ when T_s is used to

estimate the variance. The notion is that, in large samples, b_m can be replaced with $\bar{v}_m$. It is difficult to assess, however, when a large sample assumption is reasonable, because it also depends on the specific analysis being undertaken on the synthetic data, that is, it depends on the analysis's sufficient statistic(s).

The estimators T_p and T_s assume that $n_{\text{syn}} = n$ (or that n_{syn} is constant across the m synthetic data sets in the case of T_s). When using count models as opposed to multinomial models, n_{syn} is stochastic and this assumption is violated. However, in a simulation study unreported here, the effect of varying n_{syn} was found to have a negligible effect on the validity of inferences, for example, confidence intervals still achieved the nominal coverage. Nevertheless, in some cases, new estimators may be required; such estimators may introduce weights $w_1 \ldots, w_m$ that relate to $n_{\text{syn}}^{(1)}, \ldots, n_{\text{syn}}^{(m)}$, the sample sizes of the m synthetic data sets.

Averaging the $m > 1$ data sets before analysing them When faced with multiple categorical data sets, analysts (and attackers) may either pool or average the data sets *before* analysing them. This is feasible only with contingency tables, as they have the same structure across the $m > 1$ data sets. There are several advantages to doing so. Firstly, it means that analysts only have to undertake their analyses once rather than multiple times, thus leading to reduced computational time. Note, although averaging leads to non-integer "counts", standard software such as the `glm` function in R can typically cope with this and still allow models to be fit. Secondly, model-fitting in aggregated data is often hampered by the presence of zero counts, but either averaging or pooling reduces the proportion of zero counts, since it only takes one non-zero across the $m > 1$ data sets to produce a non-zero when averaged or pooled.

When the NBI is used, for a given original count $f_i = N$ $(i = 1, \ldots, K)$, the corresponding mean synthetic cell count $\bar{f}_i^{\text{syn}}$ has mean and variance,

$$E(\bar{f}_i^{\text{syn}}) = N \quad \text{and} \quad \text{Var}(\bar{f}_i^{\text{syn}}) = \frac{1}{m}\big(N + \sigma N^2\big), \tag{1}$$

as the synthetic data sets are independent.

Thus, for a given original count, the variance of the corresponding mean synthetic count is inversely proportional to m, and linearly related to σ. This means that the minimum obtainable variance when σ alone is tuned - which is achieved as $\sigma \to 0$ and the NBI tends towards its limiting distribution, the Poisson - is N/m. On the other hand, increasing m can essentially take the variance to zero. If m is too large, though, the original counts are simply returned when averaged, which, of course, renders the synthesis worthless. This, perhaps, suggests the suitability of m as a tuning parameter in cases where the original counts are large and relatively low risk, such that a relatively small variance suffices.

4 Introducing the $\tau_3(k,d)$ and $\tau_4(k,d)$ metrics

When multiple synthetic data sets are generated and the mean synthetic count calculated - which is no longer always an integer - it becomes more suitable to consider the proportion of synthetic counts *within a certain distance of* original counts of k. To allow this, the metrics $\tau_3(k)$ and $\tau_4(k)$ can be extended to $\tau_3(k,d)$ and $\tau_4(k,d)$, respectively:

$$\tau_3(k,d) \coloneqq p(|f^{\text{syn}} - k| \leq d \mid f = k), \quad \tau_4(k,d) \coloneqq p(f = k \mid |f^{\text{syn}} - k| \leq d).$$

The metric $\tau_3(k,d)$ is the probability that a cell count of size k in the original data is synthesized to within d of k; and $\tau_4(k,d)$ is the probability that a cell count within d of k in the synthetic data originated from a cell of k. Unlike k, $d > 0$ does not need to be an integer. By extending the $\tau_1(k)$ metric, such that $\tau_1(k,d)$ is the proportion of synthetic counts within d of k, it follows that $\tau_3(k,d)\tau_2(k) = \tau_4(k,d)\tau_1(k,d)$.

The $\tau_3(k)$ and $\tau_4(k)$ metrics are then special cases of $\tau_3(k,d)$ and $\tau_4(k,d)$, respectively (the case where $d = 0$). For small k, these $\tau(k,d)$ metrics are intended primarily as risk metrics, because they are dealing with uniques or near uniques. However, when d is reasonably large, $\tau_3(k,d)$ and $\tau_4(k,d)$ are, perhaps, better viewed as utility metrics, because they are dealing with the proportion of uniques that are synthesized to much larger counts (which impacts utility).

When $m > 1$ is sufficiently large, tractable expressions for the $\tau_3(k,d)$ and $\tau_4(k,d)$ metrics can be obtained via the Central Limit Theorem (CLT), as the distribution of each mean synthetic count can be approximated by a normal distribution, with mean and variance as given in (1). That is, given an original count $f_i = N$ $(i = 1, \ldots, K)$, when m is large, the distribution of the corresponding mean synthetic cell count $\bar{f}_i^{\text{syn}}$ is given as:

$$\bar{f}_i^{\text{syn}} \mid f_i = N, \sigma, m \sim \text{Normal}(N, (N + \sigma N^2)/m).$$

This can be used to approximate $\tau_3(k,d)$ and $\tau_4(k,d)$:

$$\begin{aligned}
\tau_3(k,d) &= p(|\bar{f}^{\text{syn}} - k| \leq d \mid f = k), \\
&= p(\bar{f}^{\text{syn}} < k + d \mid f = k) - p(\bar{f}^{\text{syn}} < k - d \mid f = k), \\
&= \Phi\left(\frac{(k+d) - k}{\sqrt{(k + \sigma k^2)/m}}\right) - \Phi\left(\frac{(k-d) - k}{\sqrt{(k + \sigma k^2)/m}}\right) \\
&= 2\Phi\left(\frac{d}{\sqrt{(k + \sigma k^2)/m}}\right) - 1, \qquad (2)
\end{aligned}$$

$$\begin{aligned}\tau_4(k,d) &= p(f=k \mid |\bar{f}^{\text{syn}}-k| \leq d) \\ &= \frac{\tau_3(k,d)\cdot\tau_2(k)}{\sum_{i=0}^{\infty} p(|f^{\text{syn}}-k| \leq d \mid f=i)\cdot p(f=i)} \\ &= \frac{\left[2\Phi\left(d\Big/\sqrt{(k+\sigma k^2)/m}\right)-1\right]\cdot\tau_2(k)}{\sum\limits_{i=1}^{\infty}\left[\Phi\left((k+d-i)\Big/\sqrt{(i+\sigma i^2)/m}\right)-\Phi\left((k-d-i)\Big/\sqrt{(i+\sigma i^2)/m}\right)\right]\cdot\tau_2(i)}\end{aligned} \tag{3}$$

where Φ is which is used to denote the cumulative distribution function (CDF) of the standard normal distribution.

5 Empirical Study

The data set synthesized here was constructed with the intention of being used as a substitute to the English School Census, an administrative database held by the Department for Education (DfE). It was constructed using publicly available data sources such as English School Census published data and 2011 census output tables. The data - along with a more detailed description of its origin - is available at [1]. While the data is constructed from public sources, it shares relevant features present in large administrative databases that serve to illustrate risk and utility in synthetic data and, specifically, the role that m plays in relation to the risk-utility trade-off. The framework developed here could be equally applied to any categorical data set.

The data comprises 8.2×10^6 individuals observed over $p = 5$ categorical variables. The local authority variable has the greatest number of categories with 326; while sex has the fewest with 4. When aggregated, the resulting contingency table has $K = 3.5 \times 10^6$ cells, 90% of which are unobserved, that is, have a count of zero.

The function `rNBI` from the R package **gamlss.dist** [16] was used to generate multiple synthetic data sets using the (σ, α)-synthesis mechanism described in Section 2. This was done for a range of σ, 0, 0.1, 0.5, 2 and 10, and 50 synthetic data sets were generated for each. This allowed comparisons to be drawn for a range of m, for example, taking the first five data sets gives $m = 5$, taking the first ten gives $m = 10$, etc.

5.1 Measuring risk

Evaluating risk in synthetic data, particularly in synthetic categorical data, is not always straightforward. Attempting to estimate the risk of re-identification [12] is not possible, because the ability to link records is lost when a microdata set is aggregated, synthesized and disaggregated back to microdata again.

The $\tau_3(1,d)$ and $\tau_4(1,d)$ metrics (that is, setting $k = 1$), introduced in Section 4, were used as risk metrics. Figure 1 in the Appendix shows that either increasing m or decreasing σ increases $\tau_3(1,d)$ and $\tau_4(1,d)$ and hence risk. There is an initial fall in the $\tau_3(1, 0.1)$ curves as m increases initially, suggesting lower

not higher risk. However, this is just owing to the small d: for example, when $d = 0.1$, the only way to obtain a mean synthetic count within 0.1 of k when, say $m = 5$, is by obtaining a one in each of the five synthetic data sets, compared to just once when $m = 1$.

When m is large, the $\tau_3(k, d)$ and $\tau_4(k, d)$ metrics can be approximated analytically through (2), which relies on the CLT. There is uncertainty in both the empirical values (owing to simulation error) and the analytical values (owing to the normal approximation), though the divergences between the empirical and analytical values are small.

In general, then, increasing m or decreasing σ increases risk. This is also shown visually in Figure 2 (Appendix), which demonstrates how m and σ can be used in tandem to adjust risk. Here, $\tau_3(1, 0.1)$ is used as the z-axis (risk) but any $\tau_3(k, d)$ or $\tau_4(k, d)$ would give similar results.

5.2 Measuring utility

As saturated models are used, increasing m (for a given σ) causes the mean synthetic counts to tend towards the original counts. This can be seen in the Hellinger and Euclidean distances given in Figure 3 (Appendix), which show an improvement in general utility when either increasing m or reducing σ.

These measures are equally relevant to risk, too, hence Figure 3 reiterates that risk increases with m. It is fairly trivial, however, that reducing simulation error increases risk and utility. It is more useful to gain an insight into the *rate* at which risk and utility increase with m, that is, the shape of the curves. For example, Figure 3, shows that increasing m has greater effect when $\sigma = 1$ than when $\sigma = 0.1$.

The utility of synthetic data can also be assessed for specific analyses by, for example, comparing regression coefficient estimates obtained from a model fit to both the observed and synthetic data. While such measures only assess the synthetic data's ability to support a particular analysis, they nevertheless can be a useful indicator to, for example, the required m needed to attain a satisfactory level of utility.

Here, the estimand of interest is the slope parameter from the logistic regression of age Y (aged $\leq 9 = 0$, $\geq 10 = 1$) on language X. A subset of the data were used, as just two of the language variable's seven categories were considered, while the age variable was dichotomised. When estimated from the original data, β_1 - which is a log marginal odds ratio - was equal to -0.0075 with a 95% confidence interval of (-0.0151, -0.0001). Note that, in order to estimate this, it was assumed that the original data constituted a simple random sample drawn from a much larger population. It is hugely doubtful whether such an assumption would be reasonable in practice, but the purpose here was just to evaluate the ability of the synthetic data to produce similar conclusions to the original data.

The analysis was undertaken in the two ways described in Section 3. Firstly, the $m > 1$ synthetic data sets were analysed separately and variance estimates were obtained through the estimator T_p. Secondly, the $m > 1$ synthetic data sets

were pooled into one data set prior to the analysis and variance estimates were obtained through the estimator T_s.

As can be seen in Figure 4, the estimates from T_p were noticeably larger than those from T_s, for small m. This was worrying for the validity of T_s - and the confidence intervals subsequently computed using T_s - especially since the sampling distribution of T_p was not approximated by a normal distribution, but by a t-distribution with ν_p degrees of freedom, thus widening confidence intervals further. This suggests that the large sample approximation that T_s relies on was not reasonable in this case.

The confidence interval computed from the original data set was compared with the confidence intervals computed from the synthetic data sets via the confidence interval overlap metric [5, 15]. This metric is a composite measure that takes into account both the length and the accuracy of the synthetic data confidence interval. Yet whether these factors are weighted appropriately is open to debate. Valid confidence intervals estimated from synthetic data, that is, confidence intervals that achieve the nominal coverage, are longer than the corresponding confidence intervals estimated from the original data, because synthetic data estimates are subject to the uncertainty present in the original data estimates, plus have additional uncertainty from synthesis. However, a synthetic data confidence interval, say, one that is $x\%$ narrower than the original data confidence interval (hence clearly invalid) would yield roughly the same overlap as, say, a confidence interval that is $x\%$ wider. Moreover, either an infinitely wide or infinitely small synthetic data confidence interval would achieve an overlap of 0.5.

The confidence interval overlap results are presented in Table 1 in the Appendix. The top frame gives the overlap values from when the data sets are analysed separately, and the bottom frame gives the results from when the data sets are pooled. It can be seen that increasing m broadly results in an increase in the overlap; and that the overlap tends towards 1 as the original and synthetic data confidence intervals converge. The confidence intervals computed using T_s are less robust as those using T_p, which is evident in the zero overlap when $m = 20$ and $\sigma = 10$. This is because, unlike the variance estimator T_p, T_s only considers the within-synthesis variability $\bar{v}_m$, not the between-synthesis variability b_m.

5.3 Tuning m and σ in relation to the risk-utility trade-off

The plots in Figure 5 (Appendix) show how m and σ can be tuned in tandem to produce synthetic data sets that sit favourably within the risk-utility trade-off. These trade-off plots, though, depend on the metrics used to measure risk and utility. Here, risk was measured by either $\tau_4(1, 0.5)$ or $\tau_4(1, 0.75)$, and utility by either confidence interval overlap (using T_p) or Hellinger distance. The Hellinger distances were standardised onto the interval of [0,1] (by dividing by the largest Hellinger distance observed and then subtracting from 1, so that 1 and 0 represent maximum and minimum utility, respectively).

It is possible to strictly dominate synthetic data sets over others, that is, obtain lower risk *and* greater utility values. For example, looking at the top-left

plot, synthetic data sets generated with $m = 50,\ \sigma = 2$ have higher risk but lower utility than when $m = 20,\ \sigma = 0.5$. These visual trade-offs are plotted using the empirical results, so are subject to variation from simulation; the confidence interval overlap values, in particular, can be volatile, especially when σ is large.

The intention is that the synthesizer produces such plots before releasing the data. Furthermore, as many metrics can be expressed analytically when using saturated models, they can be produced before the synthetic data is even generated.

6 Discussion

The setting of the synthesis mechanism's tuning parameters is a policy decision, and therefore is subjective. The general notion is that the synthesizer decides on an acceptable level of risk and maximises utility based on this; a larger m would necessitate a larger σ to maintain a given level of risk. As many metrics can be expressed as functions of the synthesis mechanism's tuning parameters, these functions' partial derivatives may be useful to determine the *rate* at which risk and utility change; for example, there may be a point where any further increases in m lead to a disproportionately small improvement in utility.

In addition to m, the synthesizer could also increase or decrease $E(n_{\text{syn}})$, the expected sample size of each synthetic data set. A single synthetic data set ($m = 1$) with $E(n_{\text{syn}}) = n$ contains roughly the same number of records as two synthetic data sets ($m = 2$) each with $E(n_{\text{syn}}) = n/2$. To generate a synthetic data set with an expected sample size of $n/2$, the synthesizer simply takes draws from NBI distributions with means exactly half of what they were previously. Reducing $E(n_{\text{syn}})$ should reduce risk, as fewer records are released, but inevitably reduces utility, too; once again, it calls for an evaluation with respect to the risk-utility trade-off.

Moreover, there are further tuning parameters that could be incorporated into this synthesis mechanism. One way would be to use a three-parameter distribution. When using a two-parameter count distribution, the synthesizer can increase the variance but cannot control how the variability manifests itself. The use of a three-parameter count distribution would allow the synthesizer to control the skewness, that is, they could change the shape of the distribution for a given mean and variance.

There are, of course, disadvantages to generating $m > 1$ synthetic data sets with the most obvious being the increased risk. Nevertheless, the potential benefits warrant further exploration, especially in relation to the risk-utility trade-off: does the gain in utility outweigh the increase in risk?

Organisations are taking a greater interest in making data - such as administrative data - available to researchers, by producing their own synthetic data. For this to be successful, organisations need to guarantee the protection of individuals' personal data - which, as more data becomes publicly available, becomes ever more challenging - while also producing data that are useful for analysts. Therefore, there needs to be scope to fine tune the risk and utility of

synthetic data effectively, and integrating m as a tuning parameter into this *a priori* framework helps to achieve this.

Acknowledgements

This work was supported by the Economic and Social Research Council (ESRC) via the funding of a doctoral studentship.

Appendix

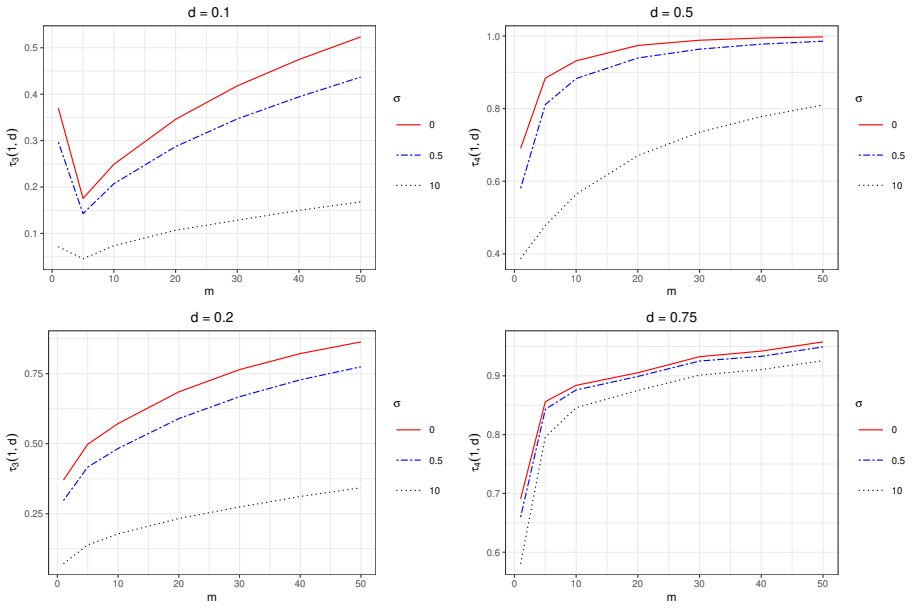

Fig. 1. The left hand plots give the empirical values of $\tau_3(1, d)$ for $d = 0.1$ and 0.2; the right hand plots give the empirical values of $\tau_4(1, d)$ for $d = 0.5$ and 0.75.

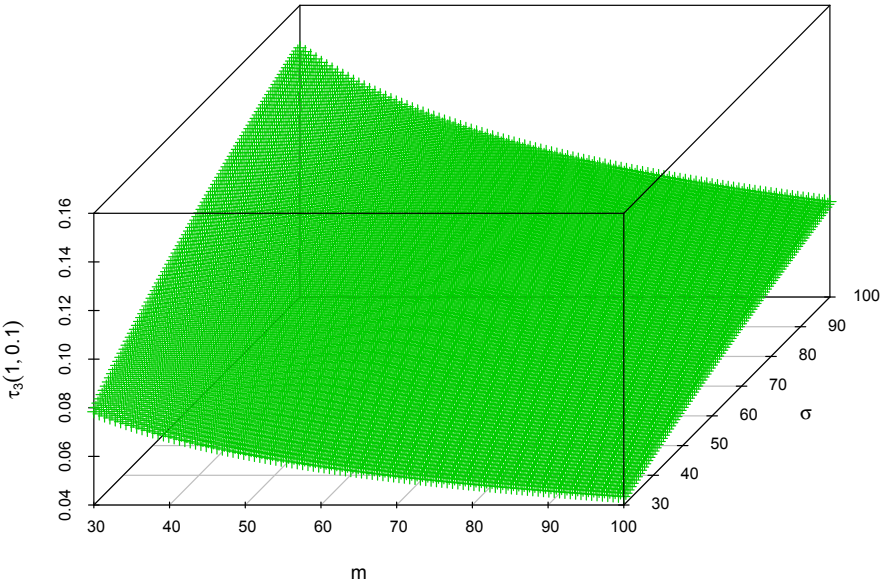

Fig. 2. The expected $\tau_3(1, 0.1)$ values for m and σ greater than 30.

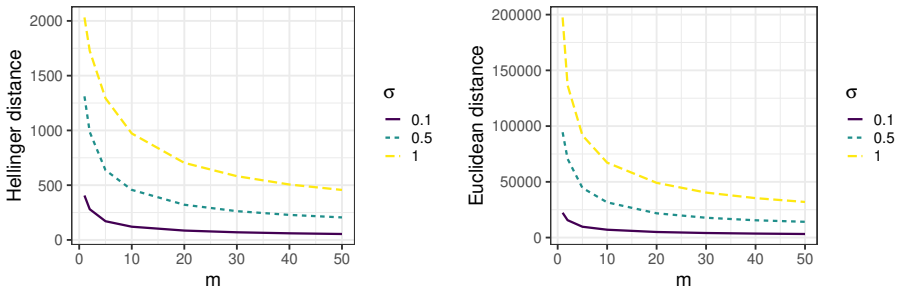

Fig. 3. The Hellinger and Euclidean distances as m increases, for various values of σ. These plots have been created using the cell counts rather than the cell probabilities, though the two are proportional.

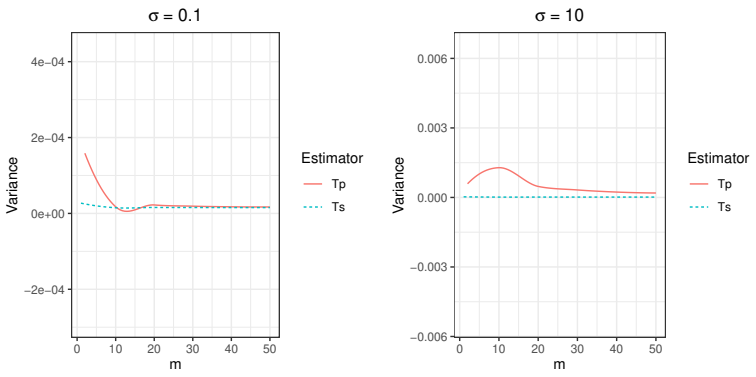

Fig. 4. The values of the estimators T_p and T_s. For small m, T_p is larger than T_s, before converging for larger m. The estimator T_s remains fairly constant across m.

Table 1. The confidence interval overlap results from when: (i) the data sets were analysed separately and T_p was used to estimate confidence intervals; and (ii) the data sets were pooled and T_s was used to estimate confidence intervals.

	$m = 2$	$m = 5$	$m = 10$	$m = 20$	$m = 30$	$m = 40$	$m = 50$
The overlap when the data sets were analysed separately and T_p used							
$\sigma = 0$	0.883	0.901	0.950	0.992	0.990	0.994	0.983
$\sigma = 0.1$	0.533	0.692	0.822	0.898	0.913	0.925	0.917
$\sigma = 0.5$	0.536	0.635	0.778	0.843	0.878	0.909	0.923
$\sigma = 2$	0.000	0.587	0.667	0.726	0.716	0.742	0.780
$\sigma = 10$	0.522	0.535	0.554	0.583	0.604	0.623	0.638
The overlap when the data sets were pooled and T_s used							
$\sigma = 0$	0.881	0.905	0.951	0.988	0.990	0.994	0.983
$\sigma = 0.1$	0.700	0.317	0.802	0.942	0.904	0.920	0.915
$\sigma = 0.5$	0.221	0.344	0.653	0.789	0.864	0.915	0.967
$\sigma = 2$	0.020	0.436	0.856	0.775	0.825	0.809	0.906
$\sigma = 10$	0.000	0.664	0.454	0.000	0.078	0.258	0.465

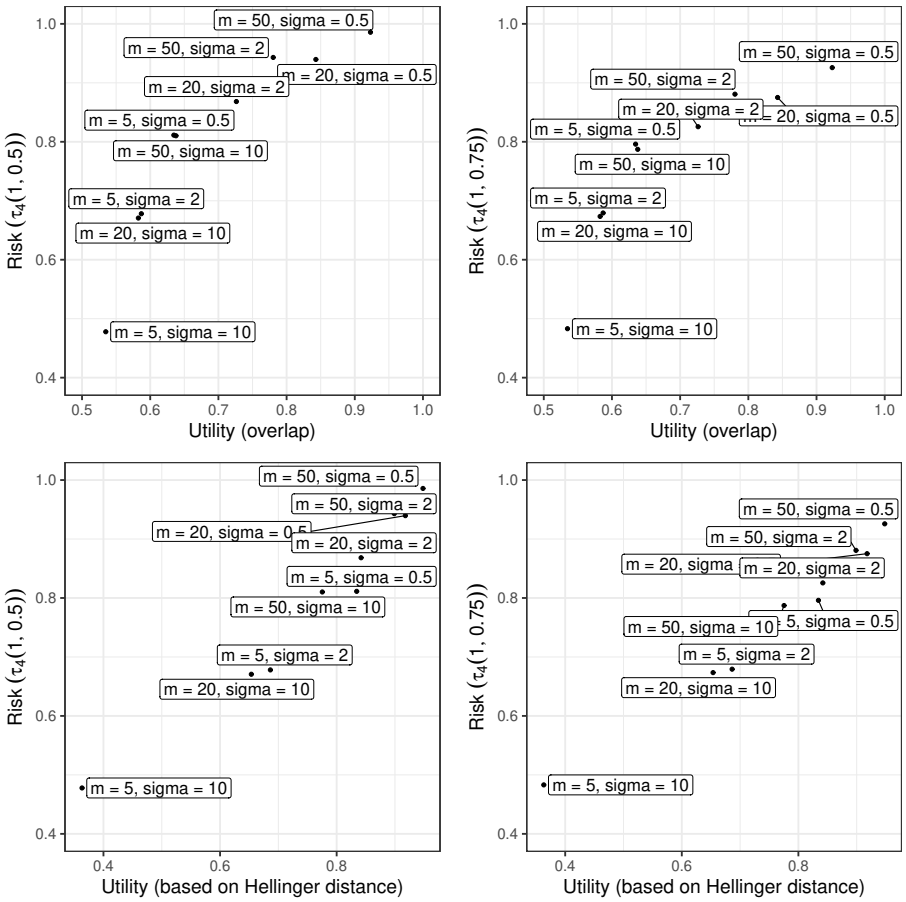

Fig. 5. Risk-utility trade-off plots to show where various synthetic data sets are located with respect to the risk-utility trade-off. The optimal position in each plot - that is, the lowest risk and the highest utility - is the bottom right corner. To measure risk, the metrics $\tau_4(1, 0.5)$ and $\tau_4(1, 0.75)$ were used. To measure utility, the confidence interval overlap and Hellinger distance were used.

References

1. Blanchard, S., Jackson, J.E., Mitra, R., Francis, B.J., Dove, I.: A constructed English School Census substitute (2022). https://doi.org/10.17635/lancaster/researchdata/533
2. Drechsler, J.: Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation, vol. 201. Springer Science & Business Media (2011)
3. Elliot, M., O'Hara, K., Raab, C., O'Keefe, C.M., Mackey, E., Dibben, C., Gowans, H., Purdam, K., McCullagh, K.: Functional anonymisation: Personal data and the data environment. Computer Law & Security Review **34**(2), 204–221 (2018). https://doi.org/https://doi.org/10.1016/j.clsr.2018.02.001, `https://www.sciencedirect.com/science/article/pii/S0267364918300116`
4. Jackson, J.E., Mitra, R., Francis, B.J., Dove, I.: Using saturated count models for user-friendly synthesis of large confidential administrative databases. Forthcoming in Journal of the Royal Statistical Society: Series A (Statistics in Society) (2022)
5. Karr, A.F., Kohnen, C.N., Oganian, A., Reiter, J.P., Sanil, A.P.: A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality. The American Statistician **60**(3), 224–232 (2006)
6. Little, R.J.: Statistical Analysis of Masked Data. Journal of Official Statistics **9**(2), 407–426 (1993)
7. Nowok, B., Raab, G.M., Dibben, C., et al.: synthpop: Bespoke Creation of Synthetic Data in R. Journal of Statistical Software **74**(11), 1–26 (2016)
8. Raab, G.M., Nowok, B., Dibben, C.: Practical Data Synthesis For Large Samples. Journal of Privacy and Confidentiality **7**(3), 67–97 (2016)
9. Raghunathan, T.E., Reiter, J.P., Rubin, D.B.: Multiple Imputation for Statistical Disclosure Limitation. Journal of Official Statistics **19**(1), 1–16 (2003)
10. Reiter, J.P.: Inference for partially synthetic, public use microdata sets. Survey Methodology **29**(2), 181–188 (2003)
11. Reiter, J.P.: Releasing multiply imputed, synthetic public use microdata: an illustration and empirical study. Journal of the Royal Statistical Society: Series A (Statistics in Society) **168**(1), 185–205 (2005)
12. Reiter, J.P., Mitra, R.: Estimating Risks of Identification Disclosure in Partially Synthetic Data. Journal of Privacy and Confidentiality **1**(1) (2009)
13. Rigby, R.A., Stasinopoulos, M.D., Heller, G.Z., De Bastiani, F.: Distributions for Modeling Location, Scale, and Shape: Using GAMLSS in R. CRC Press (2019)
14. Rubin, D.B.: Statistical Disclosure Limitation. Journal of Official Statistics **9**(2), 461–468 (1993)
15. Snoke, J., Raab, G.M., Nowok, B., Dibben, C., Slavkovic, A.: General and specific utility measures for synthetic data. Journal of the Royal Statistical Society: Series A (Statistics in Society) **181**(3), 663–688 (2018)
16. Stasinopoulos, D.M., Rigby, R.A.: Generalized Additive Models for Location Scale and Shape (GAMLSS) in R. Journal of Statistical Software **23**(7), 1–46 (2007)