Received: 31 March 2021 | Accepted: 20 April 2022

DOI: 10.1111/rssa.12876

ORIGINAL ARTICLE

ROYAL STATISTICAL SOCIETY
Series A Statistics in Society

Using saturated count models for user-friendly synthesis of large confidential administrative databases

James Jackson[1] | **Robin Mitra**[2] | **Brian Francis**[1] | **Iain Dove**[3]

[1]Lancaster University, Lancaster, UK

[2]Cardiff University, Cardiff, South Glamorgan, UK

[3]Office for National Statistics, Titchfield, Hampshire, UK

Correspondence
James Jackson, Lancaster University, Lancaster, UK.
Email: j.jackson3@lancaster.ac.uk

Abstract

Over the past three decades, synthetic data methods for statistical disclosure control have continually evolved, but mainly within the domain of survey data sets. There are certain characteristics of administrative databases, such as their size, which present challenges from a synthesis perspective and require special attention. This paper, through the fitting of saturated count models, presents a synthesis method that is suitable for administrative databases. It is tuned by two parameters, σ and α. The method allows large categorical data sets to be synthesized quickly and allows risk and utility metrics to be satisfied *a priori*, that is, prior to synthetic data generation. The paper explores how the flexibility afforded by two-parameter count models (the negative binomial and Poisson-inverse Gaussian) can be utilised to protect respondents'—especially uniques'—privacy in synthetic data. Finally, an empirical example is carried out through the synthesis of a database which can be viewed as a good substitute to the English School Census.

KEYWORDS

administrative data, categorical data, count models, data confidentiality, synthetic data

1 | INTRODUCTION

There is an increasing demand for data to be made available to researchers. This is coupled, however, with greater demands, both legal and ethical, on the protection of personal data. Consequently, new data sources, and innovative ways of protecting data, are required.

Administrative databases originate because organisations need to hold individuals' information for their day-to-day running. They have been largely under-explored as a potential data source, but can contain vast amounts of information, sometimes for an entire population. For this reason, administrative data are being used to enhance the UK census; and the UK's National Statistician, Professor Sir Ian Diamond, has gone further and recommended that administrative data can help to replace future censuses (HM Government, 2018); censuses are, after all, notoriously expensive.

Often administrative databases are hidden away with access limited to, for example, staff at government departments. When data are made available to researchers, it is usually via controlled environments, such as the Secure Research Service (SRS), a facility within the Office for National Statistics (ONS). To access data within the SRS, researchers have to undertake formal training, submit project applications and then, depending on the data's confidentiality, process the data in safe rooms. While this procedure is, of course, necessary, it can be time-consuming and may deter researchers. Alternatively, in some cases it is possible to release open data, which is when data are released into the public domain for anyone to access.

Protecting data confidentiality is the main priority when disseminating any individual-level data set. In the United Kingdom, the General Data Protection Regulation (GDPR) means that businesses and organisations are legally obliged to adhere to certain standards with regards to anonymisation when processing personal data (Information Commissioner's Office, 2020). Anonymisation is achieved through applying statistical disclosure control (SDC) methods; see Duncan et al. (2011), Hundepool et al. (2012) and Templ (2017) for a thorough review of such methods. The late Chris Skinner (1953–2020) to whom with Fred Smith (1934–2019) this special issue of Series A is dedicated, made several key contributions to the field of SDC: most notably his work on measuring disclosure risk in microdata (Skinner, 1992; Skinner & Elliot, 2002; Skinner & Shlomo, 2008; Skinner et al., 1994).

Particularly stringent anonymisation techniques are required for administrative data—even if the data are made available in a secure environment. Administrative data are particularly sensitive, as individuals do not explicitly supply their own information as they would when responding to a survey. The disclosure of sensitive information would adversely affect the reputations of those involved in processing the data, which may ultimately affect their ability to collect and process data.

The use of synthetic data (Little, 1993; Rubin, 1993) to protect privacy continues to attract attention. Whereas traditional methods typically either perturb or suppress the original data until a satisfactory level of protection is attained, synthetic data methods involve constructing new data sets by simulating from models fitted to the original data.

In 1993 the *Journal of Official Statistics* published a special issue on data confidentiality. Two contributions therein planted the seed for the growth of synthetic data sets. Rubin (1993) proposed to multiply impute values for those individuals in the population who were not sampled in the original data, and release simple random samples; these are now widely known as fully synthetic data sets (Raghunathan et al., 2003). Little (1993) proposed a similar idea whereby only certain

values in the data are replaced; these are now widely known as partially synthetic data sets (Reiter, 2003). The idea is that the synthetic data can replace the original data and provide analysts with similar inferences to those that would have been obtained had the analysis been performed on the original data.

As synthetic data are inherently artificial, disclosure risks are minimal. In statistical databases, two types of disclosure can be considered: *re-identification* and *attribute disclosure* (there is also *inferential disclosure*, although this is closely related to attribute disclosure). Re-identification is when an attacker de-anonymises an anonymised record, and so identifies an individual in the data. Attribute disclosure is when an attacker can precisely estimate an individual's sensitive values, without necessarily being able to identify them in the data set. Re-identification becomes meaningless in fully synthetic data because individuals in the synthetic data do not directly pertain to individuals in the original data. This is especially beneficial for administrative databases, which may hold information for an entire population, so there is no natural protection through sampling uncertainty, that is, uncertainty as to whether an individual was actually included in the original data. However, a synthesizer needs to guard against attribute disclosure because, through the synthetic data's release, an attacker can potentially deduce certain sensitive information. The correct attribution probability (CAP) metric, see Taub et al. (2018), seeks to measure the risk of attribute disclosure in synthetic data. For more about disclosure risk in microdata more generally see Duncan and Lambert (1989), or Hu (2019) for risk in synthetic data specifically.

The risk-utility trade-off, proposed by Duncan et al. (2001), is inherent in all data dissemination: high utility typically comes at the expense of high risk of disclosure. The ONS devised a spectrum (Bates et al., 2019) as a way of classifying the position of synthetic data with respect to this trade-off. At one end of the spectrum are 'structural' synthetic data sets where only the original data set's general structure is preserved, such as variable names; these could be employed as test data, that is, data on which researchers can first run their analyses to identify, for example, any issues with code, before repeating their analyses on the original data. At the other end of the spectrum are 'replica' synthetic data sets, which are designed to be analysed in place of the original data. The method introduced in this paper can be used to produce synthetic data anywhere on this spectrum.

The synthetic data literature stems from the literature for the multiple imputation of missing data. An appealing feature of multiple imputation is that the burden of imputing missing values falls on the imputer—a trained statistical modeller—rather than the analyst, who may be less well-versed in statistics. This philosophy carried over into the development of synthesis methods, which increasingly utilise complex computational techniques, necessitating specific non-trivial modelling decisions that require specialist training and often large amounts of recorded central processing unit (CPU) time to implement effectively. However, this is coupled with data-holders (many of whom would not be trained in these advanced statistical methods) taking greater interest in producing their own synthetic data and thus retaining greater control over the synthesis process. There is, therefore, growing appeal in developing synthesis methods that ease the burden on the synthesizer while still generating appropriate synthetic data that satisfy data-holders' requirements. This is a key motivating principle that underpins our proposed methodology.

The synthesis method presented in this paper provides a quick way to synthesize large categorical data sets. Overdispersed, saturated synthesis models are used to: (i) overcome constraints in model fitting, (ii) preserve relationships and (iii) allow risk and utility metrics to be satisfied in an *a priori* fashion. The method takes a drain-and-inject approach to synthesis: uncertainty from

modelling is drained away and, instead, uncertainty is injected where it is most needed, which is to protect the records at greatest risk of disclosure.

Through tuning two parameters (introduced later as σ and α), the synthesizer can immediately generate synthetic data with different levels of risk and utility. In this way, the method shares a trait with differentially private mechanisms (Dwork et al., 2006), which also preserve privacy through tuning a parameter (usually denoted by ϵ). Methods tuned by a parameter allow noise to be applied as appropriate: for example, when the privacy budget is high, the mechanism can easily be adjusted to reduce noise and hence risk (and vice versa). This has the potential to allow risk to be considered in a more formal way.

This paper shows that synthetic data sets are a viable option for safely making information held in administrative databases available to researchers, which will hopefully stimulate further interest and development. The paper is structured as follows: Section 2 reviews existing synthesis methods, with the focus on categorical variables, and considers particular challenges faced when synthesizing large administrative databases. Section 3 introduces this paper's contribution to the field: the (σ,α)-synthesis method, which uses saturated models. Section 4 presents an empirical illustration: the synthesis of a database which can be viewed as a substitute to the English Schools Census. Section 5 gives some concluding remarks.

2 | METHODS OF SYNTHETIC DATA GENERATION FOR CATEGORICAL DATA

Typically, data sets, including administrative databases, naturally take a microdata format, where the individuals form the rows and the variables the columns. Suppose a synthesizer wishes to synthesize a microdata set $Y = (Y_1, Y_2, \ldots, Y_p)$ comprising n individuals completely observed over p variables, so that the data form a $n \times p$ matrix. The first step in synthetic data generation involves modelling the joint multivariate distribution of this data Y. For categorical data, this can be carried out at either the microdata level or at the aggregated (tabular) level.

2.1 | Synthesizing microdata

Drechsler (2011) describes two broad methods for generating synthetic microdata: conditional and joint approaches. The conditional approaches model the original data through a product of conditional univariate models, that is,

$$p(Y_1, Y_2, \ldots, Y_p) = p(Y_1)\prod_{j=2}^{p} p(Y_j \mid Y_{j-1} \ldots, Y_2, Y_1). \tag{1}$$

Separate models can then be specified for each variable. This approach is flexible in the sense that it can deal with data sets that are comprised of different variable types. For example, normal linear regression can be used to model continuous variables and multinomial logistic regression models can be used to model categorical variables.

Joint modelling approaches, on the other hand, specify a multivariate model for the entire data set. For example, if all variables are continuous it may be possible to fit a multivariate normal distribution. However, this can be difficult to implement in practice, especially using parametric models.

2.1.1 | Conditional approaches

As described above, conditional approaches begin by modelling the first variable; then by modelling the second variable conditional on the first; the third conditional on the first and second; and so on, up to the pth variable, which is conditional on all other $p-1$ variables. Multinomial logistic regression models are an obvious choice for modelling categorical variables. Although not strictly a generalized linear model (GLM) owing to the multivariate response, the multinomial logistic regression model can be viewed as an extension to the (binary) logistic regression model to the case where the response has three or more categories.

When there are many regression coefficients, however, fitting these models is beyond the capabilities of the algorithms used in standard statistical software. Nevertheless, when the model's covariates are categorical, the time taken to obtain the regression coefficients' maximum likelihood estimates can be substantially reduced by utilising the Poisson-multinomial equivalence. Every multinomial model has a corresponding Poisson log-linear model; see, for instance, Lang (1996). Fitting the corresponding Poisson log-linear model allows the iterative proportional fitting (IPF) algorithm (Deming & Stephan, 1940) to be used, which provides a quick-and-easy route to obtain the model's fitted values (the expected counts). Skinner and Shlomo (2008) used IPF to fit log-linear models when estimating disclosure risk in microdata. However, a downside of IPF is that, while expected counts are obtained, regression coefficients' estimates and standard errors are not. This has implications for generating fully synthetic data as described by Rubin (1993), where parameters' estimates and standard errors are intrinsic to deriving parameters' posterior distributions.

Classification and regression trees (CART), which were developed by Breiman et al. (1984) and can be viewed as a non-parametric analogue to the GLM, were considered as a method to generate partially synthetic data by Reiter (2005). CART generates synthetic data sequentially, by growing a tree for each variable, conditional on all other variables in the data. Its appeal has increased with the R package **synthpop** (Nowok et al., 2016), for which CART is the default synthesis method.

Over time, CART logically led to the use of random forests for synthesis (Caiola & Reiter, 2010). Also developed by Breiman (2001), random forests grow multiple trees per variable. Drechsler and Reiter (2011) demonstrated the effectiveness of these non-parametric tree-based methods for synthesis, relative to parametric approaches.

2.1.2 | Joint modelling approaches

The non-parametric latent class model (NPLCM) (Dunson & Xing, 2009), which is a Dirichlet process mixture of products of multinomial (DPMPM) distributions, can be used to generate synthetic categorical data (Hu et al., 2014; Manrique-Vallier & Hu, 2018; Manrique-Vallier & Reiter, 2014). As with the latent class model given by Goodman (1974), the model assumes the existence of $F \geq 1$ latent classes and introduces a set of latent class probabilities $\pi_1, \ldots, \pi_F (\sum_{i=1}^{F} \pi_i = 1)$, where π_i is the probability that an individual belongs to latent class i. Then, within each latent class, a specific multinomial distribution can be fit, resulting in a flexible mixture model that can correspond to many different distributions. The model has a fully Bayesian specification, so Markov chain Monte Carlo (MCMC) methods are required to obtain samples from the posterior distribution. This can be carried out via the R package **NPBayesImputeCat** (Hu et al., 2021).

In addition, machine learning techniques are becoming an increasingly popular area of research in relation to synthetic data, such as the use of generative adversarial networks (GANs) (Kaloskampis, 2019).

2.2 | Synthesizing aggregated counts

When all variables are categorical, there are a finite number of possible observations that any individual can observe, which is determined by the number of category combinations across variables. This means that, without loss of information, the data can be expressed as a multi-dimensional contingency table, where counts give the number of times the combinations of categories are observed. In general, if there are p categorical variables with $l_1, \ldots, l_p$ categories, respectively, then the data can be cross-tabulated and expressed as a table with $K = l_1 \times \ldots \times l_p$ cell counts, where each count gives the number of individuals who belong to a particular cell.

The data can then be synthesized at the aggregated level, rather than the individual level, by modelling these counts.

2.2.1 | The Poisson log-linear model

The counts in the multi-way table can be modelled by a Poisson log-linear model, which assumes that the counts are independent and Poisson distributed. The model has a representation as a generalized linear model, in which it is parameterized by an intercept term, main effects and interaction effects. The interactions pertain to associations between variables; whenever an interaction is set to zero, independence is assumed between those variables. The ith synthetic cell count ($i = 1, \ldots, K$) of the multi-way table f_i^{syn}, is modelled as follows:

$$\begin{aligned} & f_i^{\text{syn}} \mid \beta \sim \text{Poisson}(\mu_i) \\ \text{with} \quad & \log(\mu_i) = X_i \beta. \end{aligned} \tag{2}$$

Thus the mean of f_i^{syn}, denoted by μ_i, is determined by β, the vector of log-linear model parameters (X is the design matrix).

The synthesizer must decide which interactions to include. This affects which relationships are preserved in the synthetic data. For example, if an all two-way interaction model is fitted, then relationships between all pairs of variables would be preserved, but higher order interactions—more complex relationships—would be lost. At the extreme, the saturated log-linear model includes all interactions and, as a result, each cell count in the multi-way table has its own parameter; see Agresti (2013). While the saturated model has little value when used for inference or prediction, including all interactions does ensure that all associations are preserved in the synthetic data.

The model's minimal sufficient statistics are the observed marginal tables for the highest order terms included in the model, for example, in the all two-way interaction model, all the observed two-way marginal tables. Practically, this means that the synthesizer does not require access to the full original table when synthesizing the data in this way.

As mentioned earlier, expected counts from the fitted model can be obtained relatively quickly via the IPF algorithm. The `syn.ipf` function in the R package **synthpop** (Nowok et al., 2016) implements IPF, allowing the user this choice of synthesis method.

2.2.2 | Hierarchical Poisson log-linear model

A hierarchical Poisson log-linear model can also be used to synthesize the counts in the multi-way table, as proposed by Graham and Penny (2007). The ith synthetic cell count ($i = 1, \ldots, K$) of the multi-way table f_i^{syn}, is modelled as follows:

$$\begin{aligned} f_i^{\text{syn}} \mid \lambda_i &\sim \text{Poisson}(\lambda_i) \\ \lambda_i \mid \beta, \xi &\sim \text{Gamma}(\xi, \xi/\mu_i), \text{with } \log(\mu_i) = X_i\beta. \end{aligned} \tag{3}$$

The mean of f_i^{syn}, denoted by λ_i, is now assumed to be Gamma distributed; the mean of which, in turn, is determined by β, the vector of log-linear model parameters. The parameter ξ affects the variance. The marginal distribution of f_i^{syn}, found by integrating over λ_i, is the negative binomial distribution.

There are several sources of uncertainty in this model, including model uncertainty, as decisions are made as to which interactions to include in β. Moreover, Graham and Penny (2007) generate fully synthetic data using the Bayesian posterior predictive distribution, so the β are also assumed to be stochastic.

2.3 | Challenges faced when synthesizing administrative data

The statistical challenges of dealing with administrative data are well documented; see Hand (2018). And there are further challenges when synthesizing administrative data, in addition to the usual challenges faced when synthesizing any data set, such as finding the optimal balance between risk and utility.

Henceforth it is assumed that all variables in an administrative database are categorical. This assumption is not as strong as it might first appear. Continuous variables are often subject to rounding, for example, ages given as integers; or they can be categorized by the synthesizer, for example, ages can be *converted* to integers. Besides, in any data set (n individuals), the continuous variables take a finite number of values (maximum of n values). This assumption allows the data to be expressed as a multi-way table.

2.3.1 | Large data sets

In microdata format, administrative databases are typically much larger than traditional survey data sets, especially in terms of the number of records n (the number of rows in the microdata).

For categorical data sets expressed as a multi-way table, the data's size is not governed by rows and columns—but by cells. Neither the number of individuals n, nor the number of variables p, affects the table's dimensions. Instead, the dimensions are determined by the number of categories across the variables. As such, when synthesizing the data at the aggregated level, large n can be beneficial, because it likely reduces the number of cells with zero counts (zero cells), and therefore relieves some of the problems caused by zero cells, discussed in Section 3.2.

Administrative databases can include categorical variables with many categories. When fitting models such as the Poisson log-linear model, this increases the number of parameters to estimate, thus causing the computational time to increase. It may be infeasible to fit models in such cases. The issue with computational time also extends to post-synthesis evaluations that are essential in

examining the synthetic data's risk and utility. For example, the Bayesian estimation of disclosure risk given in Reiter et al. (2014) is computationally intensive even for relatively small survey data sets, as it involves continually re-fitting the synthesis model for every individual in the data, and ways such as importance sampling have already been incorporated to save time.

2.3.2 | Random and structural zeros

The presence of large categorical variables inevitably means that multi-way tables are sparse, that is, they have a high proportion of cells with zero counts. The zero cells can be said to consist of two sorts, *random zeros* and *structural zeros*. Random zeros are zero cells that arise through random chance: an individual with a given set of characteristics could have occurred but did not in the observed data. Structural zeros (as discussed in Bishop et al., 1975) are zeros that arise because a given set of characteristics is not possible, for example, a child aged three attending a secondary school. When modelling contingency tables, structural zeros are usually dealt with by either removing the offending rows from the data set (there is no need for a balanced design when analysing contingency tables), or by weighting them out by incorporating a weight variable with weights of zero.

It is desirable that post-synthesis, all structural zeros remain zero, and some random zeros are transformed into non-zero counts. Whenever random zeros are never synthesized to non-zeros, but some non-zeros are synthesized to zeros, it results in an inflated number of zeros in the synthetic data. This issue is discussed in greater detail in Section 3.2.

In many models, such as in a Poisson log-linear model that is not saturated, it is difficult to account for structural zeros because cell means are smoothed to become non-zero. There are some exceptions, including the synthesis mechanism described by Manrique-Vallier and Hu (2018), which extends the non-parametric latent class model to account for structural zeros.

2.3.3 | No defined sampling frame

Typically, administrative data are more akin to a census than a survey, and thus more akin to population data than sample data. However, the population from which the data are drawn is unlikely to be well-defined, nor are the data likely to constitute a simple random sample from this unclear population. For example, the English School Census, an administrative database held by the Department for Education, includes pupils who attend state schools, but those who attend privately funded schools are excluded.

In general, obtaining inferences from administrative data requires careful consideration, and there are further considerations for synthesis. It can restrict the type of synthetic data which can be produced, as it may not be possible to generate fully synthetic data in the sense of Raghunathan et al. (2003), which requires the generation of a synthetic population. This is difficult when the population in question is not obvious.

Moreover, risk evaluations in SDC often revolve around estimating the probability that an individual who is unique in the sample is also unique in the population (Skinner et al., 1994). These well-established notions of *sample uniqueness* and *population uniqueness* become hazy when dealing with administrative data.

3 | THE (σ, α)-SYNTHESIS MECHANISM

Modelling for the purpose of generating synthetic data holds a unique position within statistical modelling: the objective is neither inferential nor predictive. Instead, the objective is solely to obtain synthetic data that resemble the original, but where disclosure risks are sufficiently low. That is, the model itself is not of interest. The use of saturated models, as proposed here, exploits this notion, alongside the notion that synthetic data can be obtained by sufficiently diverging away from the original data. The original data itself can be viewed as having maximum utility—but also maximum risk. A synthetic data set can be generated by trading utility for disclosure protection, so that an acceptable balance between risk and utility is achieved.

Using saturated models helps to avoid the loss of relationships between variables. In multiple imputation, when the imputation model is less complex than the analyst's model subsequently fitted to the data, the analyst's model is said to be 'uncongenial' to the imputation model (Meng, 1994). The same applies to synthesis models. Therefore, over-fitting is preferable to under-fitting, and fitting saturated models is over-fitting in its most extreme.

Moreover, saturated synthesis models eliminate bias and model uncertainty. First, this means that the synthetic counts have an unbiasedness property: the expected counts in the synthetic data are equal to those in the original data. Second, it means that there are fewer sources of uncertainty: there is only one source—that from simulation. These two points mean that the synthesizer has greater control of the synthesis mechanism, because certain properties of the synthetic data can be derived analytically, rather than needing to be found empirically. As synthetic data generation is an iterative process—data sets are generated, evaluated, improved upon and then regenerated—these analytical properties can improve the efficiency of the synthesis.

The `syn.catall` function in the R package **synthpop** (Nowok et al., 2016) facilitates the use of a saturated multinomial model to produce synthetic data. Here, count models are considered instead.

3.1 | Synthesis through saturated count models (introducing σ)

3.1.1 | The Poisson model

Suppose a saturated Poisson log-linear model is fitted to the original data's entire multi-way table. Then each count in the multi-way table is assumed to be independent and Poisson distributed, with mean equal to the observed count. Synthetic counts can be generated by simulating from these Poisson random variables, which adds stochastic error to the original counts and masks their true values.

The ith synthetic cell count f_i^{syn} $(i = 1, \ldots, K)$ of the multi-way table, is modelled as follows:

$$\begin{aligned} & f_i^{\text{syn}} \mid \mu_i \sim \text{Poisson}(\mu_i) \\ \text{with} \quad & \mu_i = f_i. \end{aligned} \tag{4}$$

where f_i is the corresponding count in the original data $(i = 1, \ldots, K)$. The difference with this model, compared to the model in (2), is that the ith synthetic count's mean μ_i is just equal to the original count f_i.

Properties of the synthesis mechanism relate directly to properties of the Poisson distribution. For example, the Poisson distribution's probability mass function gives the probability that an arbitrary synthetic count f^{syn} equals N_2, given that the original count f equals N_1:

$$p(f^{\text{syn}} = N_2 \,|\, f = N_1) = \frac{\exp(-N_1) N_1^{N_2}}{N_2!},$$

where N_1 and N_2 are non-negative integers.

While the Poisson distribution is degenerate when the mean is zero, practically, this does not affect the method: whenever an original count is zero, the synthetic count is also zero. Conveniently, this feature naturally accounts for structural zeros, which rightly remain zero. Random zeros, on the other hand, do need to be accounted for; a proposed solution is given in Section 3.2.

This synthesis mechanism produces completely synthesized data using the terminology of Raab et al. (2016). However, somewhat confusingly as a synthetic population is not created, the data are partially synthetic in the sense of Reiter (2003), rather than fully synthetic in the sense of Raghunathan et al. (2003); incidentally, Drechsler (2018) seeks to clear up some of the confusion surrounding the term 'fully synthetic' data sets. Finally, the synthesis is via the 'plug-in approach' (Reiter & Kinney, 2012), that is, the Bayesian posterior predictive distribution is not used: synthetic counts are simulated directly from the fitted model.

It follows, naturally, that the expected 'sample' size of the synthetic data n_{syn} is equal to n, the sample size of the original data. The value n_{syn}, which is stochastic, is the sum of the cell counts in the multi-way table—the table's grand total; and, as these counts are independent Poisson random variables whose means sum to n, n_{syn} is also a Poisson random variable with mean n. Yet it need not be the case that $\mathbb{E}[n_{\text{syn}}] = n$. As Raab et al. (2016) demonstrate, in completely synthesized data, n_{syn} can be made higher or lower than n. Rather than the Poisson, the multinomial can be fit here using the same framework, which would guarantee that $n_{\text{syn}} = n$. Although it is worth considering whether fixing n_{syn} is necessary—or even appropriate—with synthetic administrative data. Unlike a census, which has a known population total, an administrative database is unlikely to derive from a well-defined population, as discussed earlier in Section 2.3.3.

3.1.2 | Two-parameter count distributions for synthesis

The Poisson variability (variance equal to the mean), may not provide sufficient protection to at-risk records in the original data. The variability can be increased—without introducing bias—by using overdispersed count distributions in place of the Poisson. In a modelling context, these distributions, such as the negative binomial (NBI), are suitable whenever the sampling variance exceeds that which is expected from the Poisson.

The hierarchical Poisson log-linear model given in Section 2.2.2 essentially assumes the cell counts follow a NBI distribution. A saturated model can be fit here, too, by again including all interaction effects. The shape parameter (denoted by σ below) is set by the synthesizer, not least because there is insufficient degrees of freedom to estimate this parameter through maximum likelihood estimation. The notion is that the synthesizer determines, *a priori*—that is, prior to synthesis—the variability required to achieve a pre-specified desired level of privacy and then adjusts the variance accordingly.

When the NBI model is used, the ith synthetic cell count f_i^{syn} $(i = 1, \ldots, K)$ of the multi-way table, is modelled as follows:

$$\begin{aligned} f_i^{\text{syn}} \mid \mu_i, \sigma &\sim \text{NBI}(\mu_i, \sigma) \\ \text{with} \quad \mu_i &= f_i. \end{aligned}$$

As with the Poisson, the NBI's probability mass function gives the probability that an arbitrary synthetic count f^{syn} equals N_2, given the original count f equals N_1,

$$p(f^{\text{syn}} = N_2 \,|\, f = N_1, \sigma) = \frac{\Gamma(N_2 + 1/\sigma)}{\Gamma(N_2 + 1)\,\Gamma(1/\sigma)} \left(\frac{\sigma N_1}{1 + \sigma N_1}\right)^{N_2} \left(\frac{1}{1 + \sigma N_1}\right)^{1/\sigma}. \tag{5}$$

The mean and variance of f^{syn} are given as:

$$\mathbb{E}[f^{\text{syn}} \,|\, f = N_1, \sigma] = N_1 \text{ and } \text{Var}\,[f^{\text{syn}} \,|\, f = N_1, \sigma] = N_1 + \sigma N_1^2, \tag{6}$$

which shows how the parameter σ controls the variance of the model. There is an array of two-parameter count distributions that can be used here. The other that is considered in this paper is the Poisson-inverse Gaussian (PIG) distribution (see Rigby et al., 2019):

$$\begin{aligned} f_i^{\text{syn}} \mid \mu_i, \sigma &\sim \text{PIG}(\mu_i, \sigma) \\ \text{with} \quad \mu_i &= f_i, \end{aligned}$$

and, again, the probability that an arbitrary synthetic count f^{syn} equals N_2, given that the original count f equals N_1 is:

$$\begin{aligned} p(f^{\text{syn}} = N_2 | f = N_1, \sigma) &= \left(\frac{2c}{\pi}\right)^{1/2} \frac{N_1^{N_2} \exp(1/\sigma) K_{N_2 - 1/2}(c)}{(c\sigma)^{N_2} N_2!}, \\ \text{where} \quad c^2 = \frac{1}{\sigma^2} + \frac{2N_1}{\sigma} \quad &\text{and} \quad K_\lambda(t) = \frac{1}{2}\int_0^\infty x^{\lambda - 1} \exp\left\{-\frac{1}{2}t(x + x^{-1})\right\} \mathrm{d}x \end{aligned} \tag{7}$$

is the modified Bessel function of the third kind.

The parameterisations are as presented in the R package **gamlss.dist** (Stasinopoulos & Rigby, 2007). The NBI and PIG distributions are both continuous mixtures of Poisson distributions. In the NBI the mixing distribution is the Gamma; in the PIG the mixing distribution is the inverse Gaussian distribution. They have identical mean and variance functions (given in 6), however, higher moments differ. The key point from these two-parameter distributions is that there is a parameter, σ, that is set by the synthesizer.

3.2 | Dealing with zero counts through additive smoothing (introducing α)

There is a downside with using saturated models that needs addressing: as there is no smoothing, zero cells in the original data are always synthesized to zero, resulting in too many zeros in the

synthetic data. That is, there are the zeros from the original data, plus some non-zero cells that become zero through simulation. An excess of zero cells can affect the risk and utility of the synthetic data.

With regards to risk, the issue is not so much with the zero cells themselves, which are relatively low risk, but with what can be deduced from the non-zero cells. It follows that any non-zero cell in the synthetic data must have originated from a non-zero cell. So, from a non-zero cell in the synthetic data, an attacker can ascertain that *at least one* individual belonged to that same cell in the original data.

The addition of a pseudocount $\alpha > 0$ (which despite its name is not typically an integer) to all random zeros in the original data (structural zeros should remain zero) opens the possibility that zero counts are synthesized to non-zeros. For example, when the Poisson model is used and $\alpha > 0$ is added, the probability that a random zero $f = 0$ is synthesized to $f^{\text{syn}} = N_2$ is:

$$p(f^{\text{syn}} = N_2 \,|\, f = 0, \alpha) = \frac{\exp(-\alpha)\,\alpha^{N_2}}{N_2!}.$$

The `syn.catall` function in **synthpop** (Nowok et al., 2016), which, as mentioned earlier, uses a saturated multinomial to produce synthetic data, allows the synthesizer to specify a Dirichlet prior; this is analogous to the addition of a pseudocount proposed here.

3.3 | Tuning σ and α to satisfy metrics *a priori*

The upshot of this synthesis mechanism is that there are two parameters, σ and α, that are controlled by the synthesizer. These can be tuned to satisfy certain risk or utility metrics. The following τ metrics, are an example of a simple set of metrics that can not only be tuned by σ, but can also be represented analytically.

$\tau_1(k)$:= The proportion of cells of size k in the synthetic data.
$\tau_2(k)$:= The proportion of cells of size k in the original data.
$\tau_3(k)$:= The proportion of cells of size k in the original data, which remain of size k in the synthetic data.
$\tau_4(k)$:= The proportion of cells of size k in the synthetic data, which were also of size k in the original data.

Metrics τ_1, τ_2 and τ_4 are conditional on the distribution of cell sizes in the original data's multi-way table, whereas τ_3 is not. To illustrate, suppose a data set is comprised entirely of cell counts of one and that the Poisson distribution is used to generate synthetic counts. Then $\tau_3(1) = \exp(-1)$, which is given by the Poisson's probability mass function; and $\tau_4(1) = 1$, because the original data contains only ones. Now, when an original data set comprises a range of non-zero counts, then $\tau_3(1) = \exp(-1)$ remains unchanged, but $\tau_4(1) \leq 1$ because a synthetic cell count of one could have originated from any non-zero cell.

As structural zeros are not involved in the synthesis and just remain zero, when $k = 0$ these τ metrics refer to random zeros, for example, $\tau_1(0)$ is the proportion of random zeros in the synthetic data.

The expected values of these τ metrics can be derived analytically, as demonstrated below for when the Poisson model is used for synthesis.

3.3.1 | The metrics τ_1 and τ_2

The metric $\tau_1(k)$ is the proportion of cells of size k in the synthetic data, that is,

$$\tau_1(k) = p(f^{\text{syn}} = k) \quad k = 0, 1, 2, \ldots \tag{8}$$

which, by the law of total probability,

$$= \sum_{j=0}^{\infty} p(f^{\text{syn}} = k \,|\, f = j) \cdot p(f = j) = \frac{\exp(-\alpha)\alpha^k}{k!} \cdot \tau_2(0) + \sum_{j=1}^{\infty} \frac{\exp(-j)j^k}{k!} \cdot \tau_2(j),$$

where $\tau_2(k)$ is simply the proportion of cells with a count of k in the original data denoted by,

$$\tau_2(k) = p(f = k) \quad k = 0, 1, 2, \ldots . \tag{9}$$

3.3.2 | The metric τ_3

The metric $\tau_3(k)$ gives the proportion of cells of size k in the original data, which remain of size k in the synthetic data, that is,

$$\begin{aligned} \tau_3(k) &= p(f^{\text{syn}} = k \,|\, f = k) \quad k = 0, 1, 2, \ldots \\ &= \begin{cases} \exp(-\alpha)\alpha^k/k! & \text{if } k = 0 \\ \exp(-k)k^k/k! & \text{if } k \geq 1 \end{cases} \end{aligned} \tag{10}$$

3.3.3 | The metric τ_4

The metric $\tau_4(k)$ is the proportion of cells of size k in the synthetic data, which were also of size k in the original data, that is,

$$\tau_4(k) = p(f = k \,|\, f^{\text{syn}} = k) \quad k = 0, 1, 2, \ldots \tag{11}$$

The metric $\tau_4(k)$ can be expressed in terms of the other τ metrics:

$$\begin{aligned} \tau_4(k) = p(f = k \,|\, f^{\text{syn}} = k) &= \frac{p(f^{\text{syn}} = k \,|\, f = k) \cdot p(f = k)}{p(f^{\text{syn}} = k)} = \frac{\tau_3(k) \cdot \tau_2(k)}{\tau_1(k)} \\ &= \begin{cases} \exp(-\alpha)\alpha^k \cdot \tau_2(0) \Big/ \left(\exp(-\alpha)\alpha^k \cdot \tau_2(0) + \sum_{j=1}^{\infty} \exp(-j)j^k \cdot \tau_2(j) \right) & \text{if } k = 0 \\ \exp(-k)k^k \cdot \tau_2(k) \Big/ \left(\exp(-\alpha)\alpha^k \cdot \tau_2(0) + \sum_{j=1}^{\infty} \exp(-j)j^k \cdot \tau_2(j) \right) & \text{if } k \geq 1 \end{cases} \end{aligned}$$

3.3.4 | The τ metrics' link to disclosure risk

The notion of disclosure risk is different for synthetic data in tabular format, than for microdata. When microdata are aggregated and synthesized, the direct links between individuals in the original and synthetic data are lost.

Uniques are individuals who belong to a cell with a count of one, and are often considered to be most at risk of disclosure. An important value with respect to risk is $\tau_4(1)$: the proportion of uniques in the synthetic data which were also unique in the original data. This is arguably more important than $\tau_3(1)$: the proportion of uniques in the original data which are also unique in the synthetic data. This is because the former assumes knowledge of the synthetic data, which an attacker has access to; whereas the latter assumes knowledge of the original data, which an attacker cannot access.

There are two ways in which the synthesizer can reduce $\tau_4(1)$. The first way is to increase σ, which increases the variance of the synthetic counts. The second way is to increase α. Zero cells with small $\alpha > 0$ added are much more likely to be synthesized to one than to any other non-zero value. For example, when $\alpha = 0.1$ and the Poisson model is used, a zero count is exactly 20 times more likely to be synthesized to one than two, which increases the number of uniques in the synthetic data and thereby decreases $\tau_4(1)$.

3.3.5 | Tuning σ and α to adjust the expected values of the τ metrics

The notion is that the synthesizer tunes σ and α to yield synthetic data with certain properties. As an example, the synthesizer can decide, *a priori*, that they would like $\tau_1(0) = \tau_2(0)$ and $\tau_4(1) = p$ (for some p)—and they can then tune (σ, α) accordingly.

When $\alpha = 0$, the inequality $\tau_1(0) \geq \tau_2(0)$ holds, since zero cells in the synthetic data comprise all zero cells in the original data, plus those that randomly become zero through synthesis. As α increases, the expected difference between $\tau_1(0)$ and $\tau_2(0)$ narrows and the inequality would eventually reverse.

An attractive property might be for the synthetic data to have the same proportion of zero cells as the original data, that is, for $\tau_1(0) = \tau_2(0)$. Under the Poisson model this is achieved by setting:

$$\alpha^* = -\log\left\{1 - \frac{1}{\tau_2(0)}\sum_{j=1}^{\infty}\exp(-j)\cdot\tau_2(j)\right\}.$$

The short derivation is given in the supplementary material. An alternative is to choose α such that $\tau_4(1) = p$, where the synthesizer decides what $p \in [0, 1]$ is acceptable (a pre-specified level of disclosure risk). Here the value of α^* must be obtained numerically; under the Poisson model it satisfies,

$$p = \exp(-1)\cdot\tau_2(1)\Big/\left(\alpha^*\exp(-\alpha^*)\cdot\tau_2(0) + \sum_{j=1}^{\infty} j\cdot\exp(-j)\cdot\tau_2(j)\right).$$

Similar expressions can be derived for the two-parameter synthesis models, where the required value of α^* also depends on σ. These are also included in the supplementary material.

4 | EMPIRICAL STUDY

4.1 | The data

The English School Census (ESC) is an administrative database that holds information about pupils in state-funded schools. Every school term the Department for Education (DfE) requests

TABLE 1 The ESCsub's variables and their numbers of categories

Variable	Type	# Categories
Area Code/ Geography (V)	Categorical	326
Ethnicity (W)	Categorical	20
Sex (X)	Categorical	4
Age (Y)	Categorical	19
Language (Z)	Categorical	7

that all nursery, primary and secondary schools, which are fully or partly funded by the state, submit details about the school and its pupils. This is just one example of an administrative database held by a government department; other examples include, but are not limited to, the Patient Register (held by the Department of Health) and the Customer Information System (held by the Department for Work and Pensions).

For obvious reasons, access to the ESC data, as well as to other administrative databases, is highly restricted. However, in previous work conducted by the ONS, a carefully constructed data set using publicly available sources was created to be used as a substitute to the ESC, in order to develop synthesis methods for administrative data. These data were used here as the basis for generating a synthetic database.

The data were constructed using public 2011 census output tables involving various combinations of local authority, sex, age and ethnicity.[1] Language attributes from the census were also included and artificially expanded to match with categories in the ESC. In addition, school phase attributes were incorporated, some adjustments for migration were applied, and non-response and invalid categories were added to various variables, again taking publicly available information from the census.

The two variables measured at the school level were ignored for this illustration, which focused instead on the remaining five variables measured at the pupil level. Henceforth, this data set is referred to as the ESCsub where 'sub' denotes substitute. Table 1 summarises the variables present in the ESCsub illustration. The data comprise $n = 8{,}190{,}870$ pupils over $p = 5$ categorical variables, giving rise to a multi-way contingency table with $K = 326 \times 20 \times 4 \times 19 \times 7 = 3.5 \times 10^6$ cells. The data set—along with a more detailed description of its origin—is available at Blanchard et al. (2022). The breakdown of the cell counts are given in Table 2; only 333,660 (9.6%) are non-zero—so the data are sparse. There are no structural zeros.

So, while the data are in a sense simulated, this was done using real data sources and care was taken to ensure that the resulting data reflect, at the very least, the typical structure present in the ESC. As such, this was a good example to use to demonstrate our synthesis method and a similar performance is expected when the method is applied to the actual ESC, as well as other similar large categorical administrative databases. Importantly, the data were not generated from a statistical model and thus do not favour a particular synthesis method.

[1]Specifically, information from the following public sources were used to create the data: http://www.nomisweb.co.uk/census/2011; http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-user-guide/quality-and-methods/quality/quality-measures/response-and-imputation-rates/index.html; https://www.gov.uk/government/statistics/schools-pupils-and-their-characteristics-january-2014.

TABLE 2 Distribution of cell sizes in the ESCsub data

Cell count	Frequency	% of cells
0	3,134,980	90.38
1	119,917	3.46
2	51,412	1.48
3	25,952	0.75
4	19,450	0.56
5	13,076	0.38
6	10,345	0.30
7	7947	0.23
8	7077	0.20
9	5809	0.17
10	5163	0.15
≥11	67,512	1.95
Total	3,468,640	100

4.2 | The synthesis

The synthesis was carried out in R (version 3.6.3) using the methods described in this paper. The Poisson, NBI and PIG models were compared by examining how the counts in the synthetic data's multi-way table deviate from those in the original data, and by computing summaries of risk and utility. This evaluation also includes a comparison of parameter estimates obtained from a log-linear analysis, which was performed on both the original and the synthetic data.

Just $m = 1$ data set was generated for each synthesis model. The CPU times to carry this out were 0.2, 0.3 and 162 s for the Poisson, NBI and PIG models respectively. The PIG model took notably longer, although is still fast compared to other methods, such as the conditional approaches (Section 2.1.1) that synthesize the data at the microdata level.

Let V, W, X, Y and Z denote the five variables in the data, and let f_{vwxyx} denote the cell count of a particular cell in the cross-classified table corresponding to category $v \in V$, $w \in W$, $x \in X$, $y \in Y$ and $z \in Z$. A synthetic count was then drawn for this cell by,

$$f^{\text{syn}}_{vwxyz} \sim \text{Poisson}(f_{vwxyz}),$$

when the Poisson synthesis model was used; or, when either of the two-parameter distributions were used,

$$f^{\text{syn}}_{vwxyz} \sim \text{NBI}(f_{vwxyz}, \sigma) \quad \text{or} \quad f^{\text{syn}}_{vwxyz} \sim \text{PIG}(f_{vwxyz}, \sigma).$$

For the Poisson, the only parameter to be set was α, the pseudocount added to random zeros in the original data. For the two-parameter count models, there was the additional parameter σ to consider.

As mentioned previously, one of the appealing features of using these saturated synthesis models is that it allows the synthesizer to determine properties of the synthesis model *a priori*,

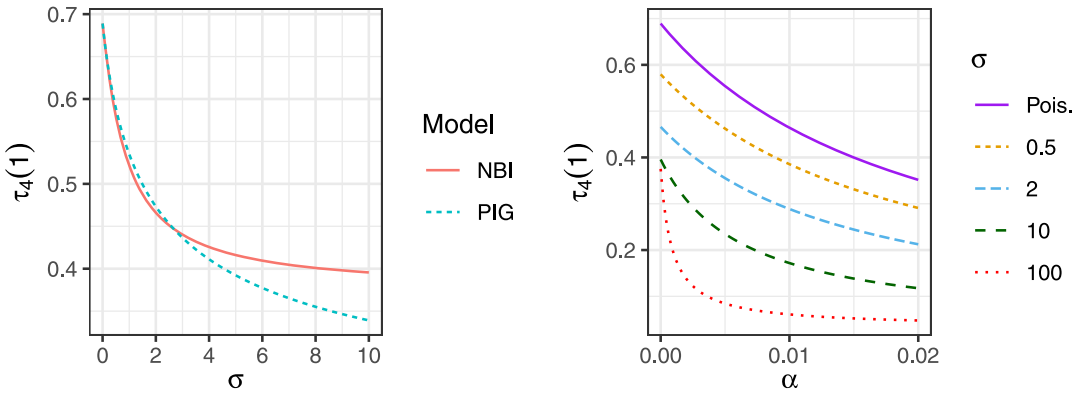

FIGURE 1 The left plot gives the risk metric $\tau_4(1)$ - the proportion of uniques in the synthetic data that were also unique in the original data - as a function of σ for when the NBI (solid line) and PIG (dashed line) models were used ($\alpha = 0$). The right plot shows how α and σ together affect $\tau_4(1)$ when the NBI is used.

thus reducing the amount of empirical evaluation necessary during the synthesis. For illustration, Figure 1 (left) compares the effect of σ on the risk metric $\tau_4(1)$ for the NBI and PIG models. For large σ, the risk levels off for the NBI but continues to fall away for the PIG. Figure 1 (right) also looks at $\tau_4(1)$, but at the combined effect of σ and α when the NBI is used. For all σ, $\tau_4(1)$ falls as α increases and $\tau_4(1)$ is always lower for the NBI than for the Poisson.

4.3 | Descriptive summaries of risk and utility

Table 3 gives the proportion of cell counts in the synthesized tables that are within p% of their original size, for different σ and α. The first block of results considers all cells, while the second block only considers non-zero cells (in the observed data). Smaller values of p can be viewed as summaries of risk while larger values of p measures of utility. To elaborate, if a large proportion of original and synthetic cell counts are very close, say, within 0.5% ($p = 0.5$) of each other, then the synthetic data could be considered to be high risk. If, on the other hand, few original and synthetic cell counts are within, for example, 50% ($p = 50$) of each other, then this is likely to indicate low utility. As an example, the 0.927 value in the top-left corner of the table means that when $\alpha = 0$ and $\sigma = 0$, 92.7% of all cell counts in the synthetic data were within 0.5% of the corresponding count in the original data. The Poisson model had the greatest utility but the greatest risk. There was little to choose between the NBI and PIG models based on these summaries. As expected, greater σ or α lead to greater divergences in original and synthetic cell sizes.

The similarities between the NBI and PIG models are also highlighted in Figures 2 and 3, which plot the synthetic (vs.) original counts and also percentage differences (between synthetic and original counts) (vs.) original counts. While the Poisson model's points ($\sigma = 0$ cases) were close to the 45° line—which indicates strong correlation between synthetic and original counts—this correlation reduces as σ increases in both the NBI and PIG models. Even a relatively small value of $\sigma = 0.01$ introduced noticeable dispersion around the 45° line. The right panels display a funnel shape, that is, percentage differences were greater for smaller counts than for larger

TABLE 3 Empirical results showing the proportion of synthetic cell counts within p% of the corresponding original counts. The table includes results for both the NBI and the PIG, for different σ and α. The upper block of results considers all original cell counts, while the lower block considers only non-zero original cells. For α=0.02, whenever a zero count was synthesized to a non-zero count, although the percentage difference was not estimable (zero denominator), it was deemed to be greater than 50% for the purpose of this table

	Proportion of synthetic cell counts within p% of the original									
	NBI					**PIG**				
p	**0.5**	**1**	**5**	**10**	**50**	**0.5**	**1**	**5**	**10**	**50**
All original cell counts										
σ										
$\alpha = 0$										
0 (Pois.)	0.927	0.927	0.931	0.935	0.967	0.927	0.927	0.931	0.935	0.967
0.1	0.924	0.924	0.926	0.928	0.961	0.925	0.925	0.926	0.928	0.961
0.5	0.920	0.920	0.920	0.922	0.946	0.921	0.921	0.921	0.923	0.949
1	0.917	0.917	0.917	0.918	0.937	0.918	0.918	0.919	0.920	0.942
2	0.914	0.914	0.914	0.914	0.928	0.916	0.916	0.916	0.917	0.935
5	0.910	0.910	0.910	0.910	0.918	0.913	0.913	0.913	0.914	0.927
10	0.907	0.907	0.907	0.908	0.912	0.911	0.911	0.911	0.912	0.921
$\alpha = 0.02$										
0 (Pois.)	0.909	0.910	0.913	0.917	0.949	0.909	0.910	0.913	0.917	0.949
0.1	0.907	0.907	0.908	0.910	0.943	0.907	0.907	0.908	0.910	0.943
0.5	0.902	0.902	0.903	0.904	0.928	0.903	0.903	0.903	0.905	0.931
1	0.899	0.899	0.900	0.901	0.920	0.901	0.901	0.901	0.902	0.924
2	0.896	0.896	0.896	0.897	0.911	0.899	0.899	0.899	0.900	0.918
5	0.893	0.893	0.893	0.893	0.901	0.896	0.896	0.896	0.897	0.909
10	0.891	0.891	0.891	0.891	0.896	0.895	0.895	0.895	0.895	0.905
Non-zero original cell counts										
σ										
$\alpha = 0$										
0 (Pois.)	0.242	0.245	0.280	0.327	0.658	0.242	0.245	0.280	0.327	0.658
0.1	0.214	0.215	0.226	0.252	0.592	0.217	0.218	0.229	0.256	0.598
0.5	0.167	0.167	0.173	0.187	0.437	0.177	0.177	0.182	0.197	0.468
1	0.136	0.136	0.140	0.150	0.347	0.153	0.153	0.157	0.167	0.395
2	0.102	0.102	0.105	0.111	0.253	0.128	0.128	0.131	0.138	0.324
5	0.059	0.059	0.061	0.064	0.145	0.097	0.097	0.099	0.104	0.238
10	0.037	0.037	0.038	0.040	0.089	0.076	0.076	0.077	0.081	0.183

(Continues)

TABLE 3 (Continued)

	Proportion of synthetic cell counts within p% of the original									
	NBI					PIG				
p	0.5	1	5	10	50	0.5	1	5	10	50
$\alpha = 0.02$										
0 (Pois.)	0.242	0.245	0.279	0.326	0.657	0.242	0.245	0.279	0.326	0.657
0.1	0.215	0.215	0.226	0.253	0.593	0.215	0.216	0.227	0.254	0.598
0.5	0.167	0.167	0.172	0.186	0.437	0.175	0.176	0.181	0.196	0.468
1	0.137	0.137	0.141	0.151	0.348	0.153	0.153	0.157	0.167	0.396
2	0.102	0.102	0.105	0.111	0.253	0.128	0.128	0.130	0.138	0.324
5	0.061	0.061	0.062	0.065	0.147	0.096	0.096	0.098	0.103	0.237
10	0.037	0.037	0.037	0.039	0.088	0.076	0.076	0.077	0.081	0.182

counts. This is an ideal profile for balancing risk and utility, as the riskiest individuals are the ones corresponding to small cell counts, and these cell counts require the most movement during synthesis. On the other hand, large counts are relatively low risk, and proportional changes to large counts will have a more significant impact on utility, thus relatively less perturbation is desired.

Table 4 presents empirical values for the τ metrics, again for varying σ and α. The expected values are known prior to synthesis, although a small difference occurs, owing to simulation noise. But, for cell sizes that are prevalent in the original data—such as zeros and ones—this error is negligible. For example, the empirical value obtained for $\tau_3(1)$ when the Poisson model was used ($\sigma = 0$, $\alpha = 0$) is 0.3674, which is almost identical to the expected value, $\exp(-1) = 0.3679$.

Table 4 also illustrates the suitability of α in reducing risk. The values for $\tau_4(1)$ are substantially lower when $\alpha = 0.02$ than when $\alpha = 0$; for example, when the Poisson model is used, $\tau_4(1)$ is 0.352 compared to 0.689. For a given α, the NBI and PIG models almost always have a lower risk than the Poisson model when considering the $\tau_3(1)$ and $\tau_4(1)$ metrics. It is particularly interesting to note varying profiles between synthesis models and these metrics. For example, if one model has a lower $\tau_3(1)$ value than another, then this is not necessarily the case when comparing the corresponding $\tau_4(1)$ value. To illustrate, consider the case when $\alpha = 0$ and $\sigma = 10$. For the NBI, the value of $\tau_3(1)$ is $0.0711 < 0.1532$ the value for the PIG. But for $\tau_4(1)$, with these same parameter values, the value under the NBI is $0.3910 > 0.3387$ the value under the PIG. The specific choice of synthesis model to use would depend on the synthesizer's range of permitted values for τ_3, and τ_4, and choosing the model that best satisfies these requirements.

4.4 | Testing specific utility through log-linear model analysis

In the synthetic data literature, specific utility (Snoke et al., 2018) is often assessed by comparing inferences, such as regression coefficients, obtained from the original and synthetic data.

The synthesizer does not know, of course, what analyses users of the synthetic data would perform. Among the variables included in the data, which are best described as demographic, there

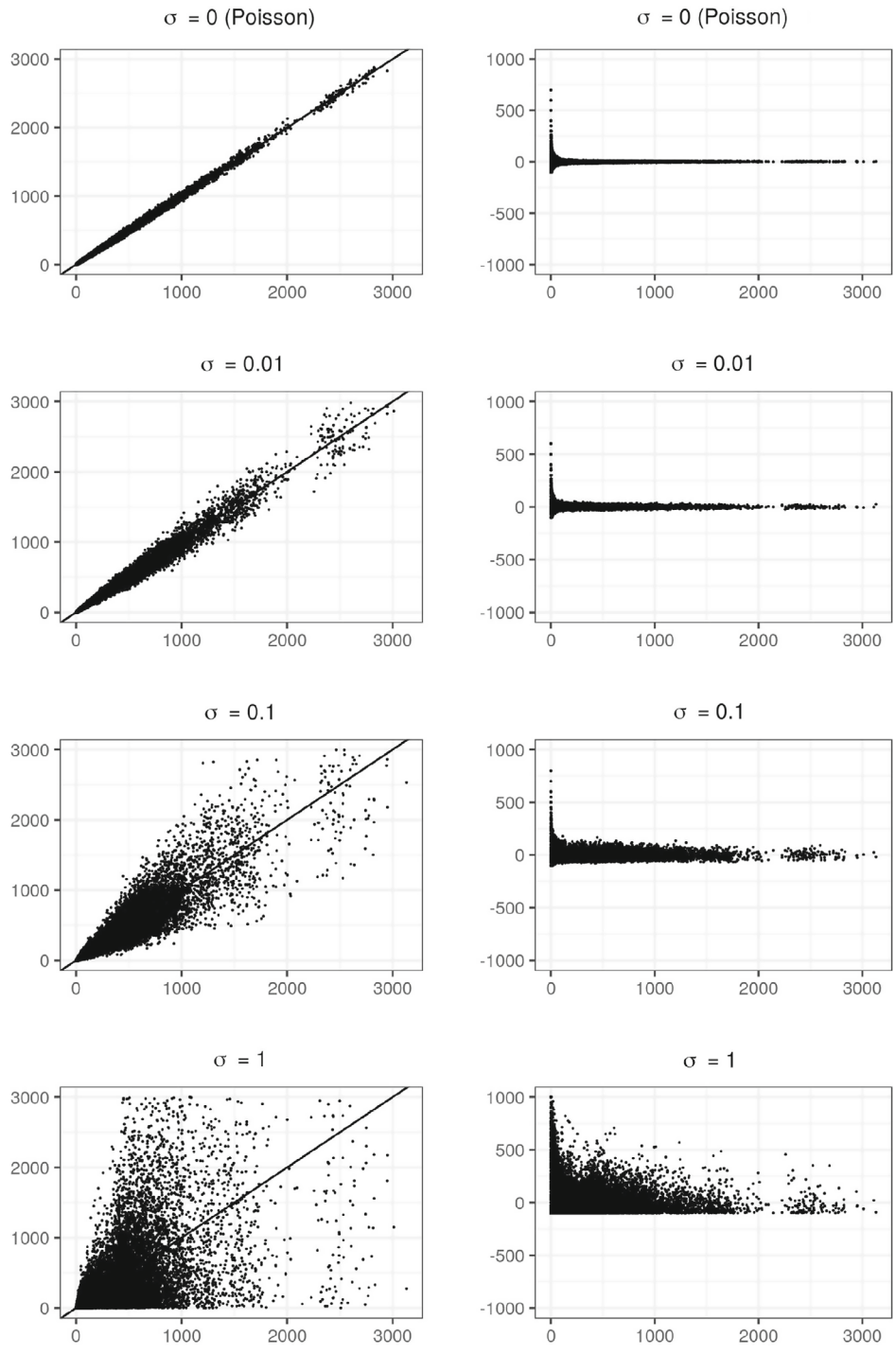

FIGURE 2 The left hand plots give the synthetic counts (vs.) the original counts for different σ when the NBI model was used for synthesis and $\alpha = 0$. The original counts of zero—which were always synthesized to zero because $\alpha = 0$—were omitted. The right hand plots give original and synthetic counts' percentage differences (vs.) the original counts. The percentage differences were calculated by: 100*(synthetic count - original count) / original count.

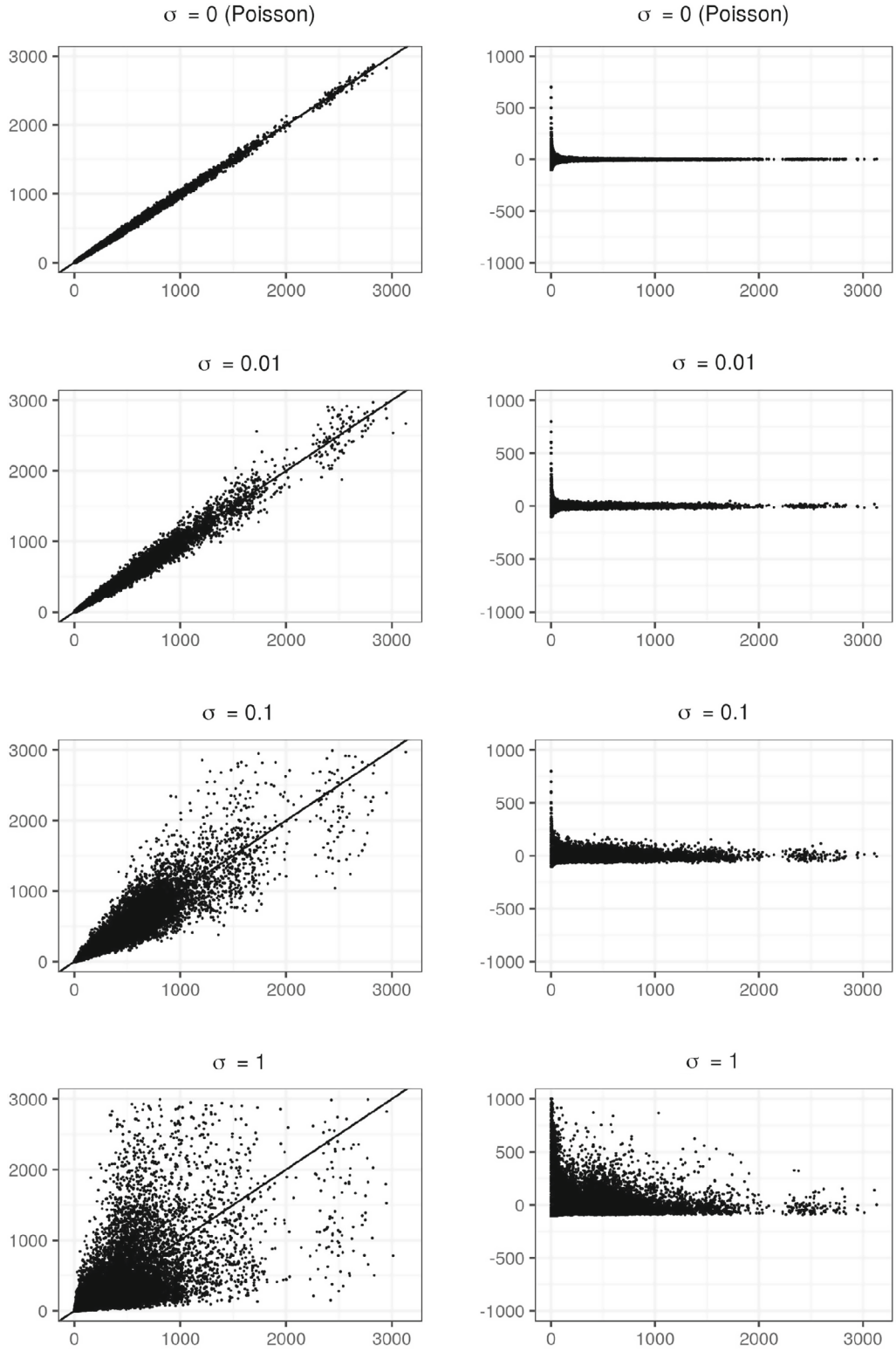

FIGURE 3 As Figure 2 but for when the PIG was used rather than the NBI.

TABLE 4 Empirical values obtained for the τ metrics for different σ and α and for the NBI and PIG

	NBI				PIG			
k	**0**	**1**	**2**	**3**	**0**	**1**	**2**	**3**
σ	$\tau_1(k)$							
$\alpha = 0$								
0 (Pois.)	0.9190	0.0184	0.0135	0.0086	0.9190	0.0184	0.0135	0.0086
0.01	0.9191	0.0184	0.0134	0.0086	0.9192	0.0184	0.0134	0.0086
0.1	0.9204	0.0183	0.0130	0.0085	0.9203	0.0184	0.0130	0.0084
0.5	0.9256	0.0177	0.0117	0.0077	0.9243	0.0181	0.0121	0.0078
1	0.9317	0.0166	0.0105	0.0068	0.9280	0.0179	0.0111	0.0072
5	0.9587	0.0098	0.0054	0.0035	0.9422	0.0167	0.0086	0.0053
10	0.9713	0.0064	0.0033	0.0022	0.9500	0.0156	0.0072	0.0042
$\alpha = 0.02$								
0 (Pois.)	0.9013	0.0359	0.0136	0.0086	0.9013	0.0359	0.0136	0.0086
0.01	0.9012	0.0362	0.0136	0.0085	0.9013	0.0362	0.0136	0.0086
0.1	0.9024	0.0360	0.0133	0.0084	0.9024	0.0361	0.0133	0.0085
0.5	0.9078	0.0352	0.0120	0.0077	0.9065	0.0357	0.0122	0.0078
1	0.9139	0.0339	0.0107	0.0069	0.9101	0.0355	0.0115	0.0072
5	0.9415	0.0259	0.0063	0.0036	0.9251	0.0330	0.0093	0.0053
10	0.9550	0.0212	0.0047	0.0024	0.9337	0.0307	0.0084	0.0045
	$\tau_2(k)$							
	0.9038	0.0346	0.0148	0.0075	0.9038	0.0346	0.0148	0.0075
	$\tau_3(k)$							
$\alpha = 0$								
0 (Pois.)	1	0.3674	0.2701	0.2231	1	0.3674	0.2701	0.2231
0.01	1	0.3676	0.2706	0.2221	1	0.3653	0.2703	0.2189
0.1	1	0.3489	0.2457	0.1976	1	0.3538	0.2484	0.1974
0.5	1	0.2964	0.1874	0.1340	1	0.3090	0.2022	0.1468
1	1	0.2499	0.1489	0.1024	1	0.2779	0.1677	0.1197
5	1	0.1144	0.0618	0.0403	1	0.1895	0.0981	0.0654
10	1	0.0724	0.0378	0.0248	1	0.1532	0.0740	0.0466
$\alpha = 0.02$								
0 (Pois.)	0.9804	0.3648	0.2695	0.2247	0.9804	0.3648	0.2695	0.2247
0.01	0.9802	0.3645	0.2695	0.2186	0.9802	0.3656	0.2669	0.2217
0.1	0.9802	0.3499	0.2450	0.1950	0.9801	0.3494	0.2466	0.1984
0.5	0.9803	0.2950	0.1876	0.1391	0.9803	0.3090	0.1981	0.1436
1	0.9804	0.2515	0.1498	0.1052	0.9803	0.2782	0.1689	0.1218
5	0.9812	0.1172	0.0620	0.0429	0.9810	0.1888	0.0959	0.0629
10	0.9819	0.0711	0.0374	0.0255	0.9817	0.1524	0.0750	0.0486

(Continues)

TABLE 4 (Continued)

	NBI				PIG			
k	0	1	2	3	0	1	2	3
	$\tau_4(k)$							
$\alpha = 0$								
0 (Pois.)	0.9835	0.6893	0.2974	0.1943	0.9835	0.6893	0.2974	0.1943
0.01	0.9834	0.6890	0.2995	0.1938	0.9833	0.6867	0.2989	0.1905
0.1	0.9820	0.6603	0.2811	0.1742	0.9821	0.6648	0.2827	0.1760
0.5	0.9764	0.5788	0.2372	0.1304	0.9778	0.5890	0.2484	0.1416
1	0.9701	0.5203	0.2108	0.1125	0.9739	0.5369	0.2232	0.1243
5	0.9427	0.4043	0.1710	0.0858	0.9593	0.3919	0.1694	0.0929
10	0.9305	0.3910	0.1677	0.0851	0.9513	0.3387	0.1521	0.0822
$\alpha = 0.02$								
0 (Pois.)	0.9831	0.3516	0.2935	0.1957	0.9831	0.3516	0.2935	0.1957
0.01	0.9829	0.3484	0.2935	0.1919	0.9830	0.3495	0.2911	0.1934
0.1	0.9817	0.3357	0.2735	0.1733	0.9817	0.3350	0.2745	0.1751
0.5	0.9759	0.2898	0.2316	0.1348	0.9774	0.2991	0.2403	0.1379
1	0.9696	0.2567	0.2066	0.1141	0.9735	0.2712	0.2168	0.1259
5	0.9419	0.1562	0.1469	0.0890	0.9584	0.1979	0.1520	0.0887
10	0.9293	0.1162	0.1172	0.0799	0.9503	0.1715	0.1320	0.0808

is no obvious response variable, so analysts may be interested in associations between variables. Therefore, a log-linear analysis was chosen as a suitable way to test the specific utility of synthetic data generated with the synthesis method described in Section 3. It is difficult to obtain parameter estimates and parameters' standard error estimates for the full five-variable data, since large amounts of memory and storage are required—the same problem faced when fitting synthesis models. To relieve some of this pressure, the all two-way interaction model was fitted to three of the data's five variables, ethnicity, age and language, resulting in 608 parameters.

The confidence interval overlap metric (Karr et al., 2006) was used to measure similarities between estimates. In order to define confidence interval overlap, let (l_o, u_o) and (l_s, u_s) denote confidence intervals for a univariate population parameter Q, obtained from the original and synthetic data, respectively; and let (l_i, u_i) denote the intersection of the two intervals, that is, $l_i = \max(l_o, l_s)$ and $u_i = \min(u_o, u_s)$. Then the confidence interval overlap I_Q is given as:

$$I_Q = \frac{1}{2}\left(\frac{u_i - l_i}{u_o - l_o} + \frac{u_i - l_i}{u_s - l_s}\right). \tag{12}$$

Thus I_Q is the mean of two ratios: the length of the confidence interval intersection divided by (i) the length of the confidence interval from the original data, and (ii) the length of the confidence interval from the synthetic data.

Combining rules are required to obtain valid parameter estimates and standard errors from synthetic data, even when just $m = 1$ synthetic data set is generated. This is because there are always two sources of uncertainty in synthetic data that need to be accounted for: the sampling

uncertainty inherent in the original data, and the uncertainty owing to synthesis. To simplify the analysis—after all, the purpose here is just to evaluate the utility of the synthetic data—the original data were assumed to constitute a simple random sample drawn from a super-population. This allowed the estimator given in Raab et al. (2016) to be used, which provides valid variance estimates for large samples when analysing synthetic data generated through the mechanism described in Section 3. When estimating a population parameter Q from $m \geq 1$ synthetic data sets, $\hat{Q}$ is found by averaging over the m data sets, and its variance is given as:

$$\widehat{Var}(\hat{Q}) = \bar{v}_m(n_{\text{syn}}/n + 1/m), \tag{13}$$

where $\bar{v}_m$ is the mean variance estimate across the m synthetic data sets, and n_{syn} and n are the 'sample' sizes of the synthetic and original data respectively. Unlike other estimators, such as the one given in Reiter (2003), this estimator allows valid variance estimates to be obtained from just $m = 1$ synthetic data set, as done here. When $m = 1$ ($\bar{v}_m = v$) and $n = n_{\text{syn}}$, the estimator in (13) simplifies to $2v$, that is, the variance estimate from the synthetic data, doubled.

Finally, in log-linear models, estimability issues can arise through the presence of zero counts in the data. This can lead to issues surrounding non-existence and non-identifiability of estimates (Fienberg & Rinaldo, 2012). But no serious model fitting issues arose in this particular example. There were some parameters included in the model with a true value of $-\infty$. For such parameters, R returned a large negative value, typically in the vicinity of -20.

4.4.1 | Results

Figures 4 and 5 present boxplots of confidence interval overlap values for the log-linear model parameters across the different synthesis models. They demonstrate how increasing σ and α causes utility to fall away. For example, irrespective of α, whenever σ=10, the median confidence interval overlap is zero. In general, a high proportion of the overlap values are equal to 1/2. This can be seen, for example, in the centre and right plots of Figure 5, where several of the upper quartiles are equal to 1/2. This, incidentally, is owing to the nature—and perhaps a criticism—of the confidence interval overlap metric (given in 12): whenever one of the confidence intervals'

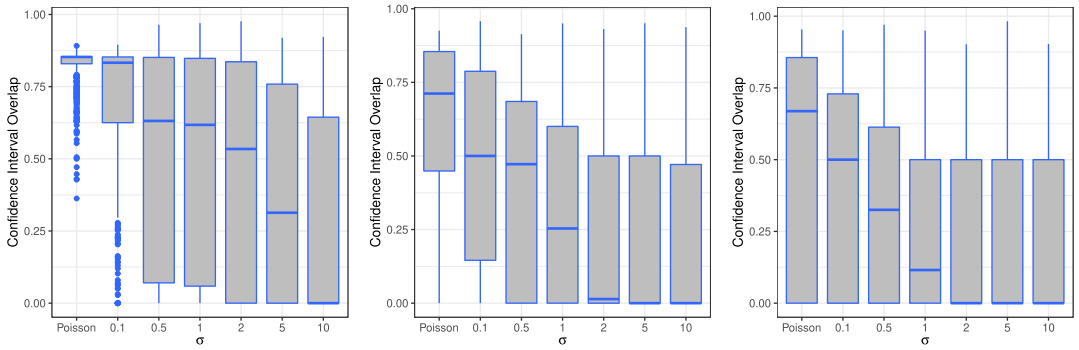

FIGURE 4 These boxplots show how σ and α affect log-linear parameters' confidence interval overlap when the NBI distribution is used for synthesis. The left frame is the case where $\alpha = 0$; the middle frame where $\alpha = 0.01$; and the right frame where $\alpha = 0.02$.

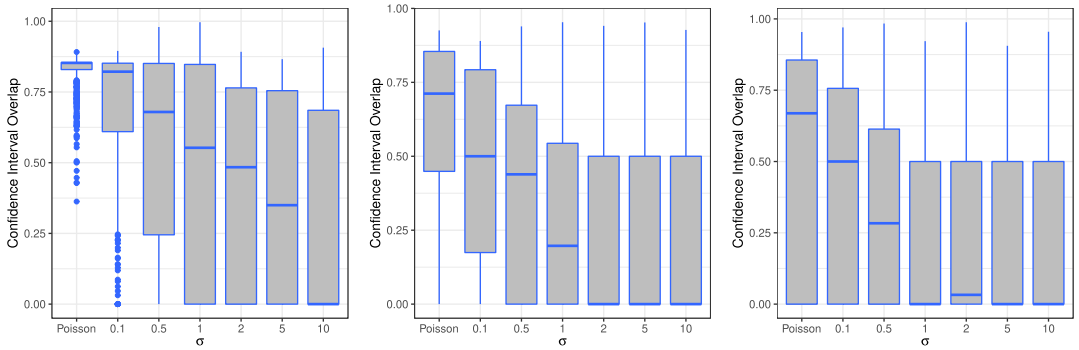

FIGURE 5 These boxplots show how σ and α affects confidence interval overlap when the PIG distribution is used for synthesis. The left frame is the case where $\alpha = 0$; the middle frame where $\alpha = 0.01$; and the right frame where $\alpha = 0.02$.

TABLE 5 How σ and α affect the trimmed mean (top and bottom 10% excluded) percentage difference between log-linear parameter estimates obtained from the observed and synthetic data. The trimmed mean was used to subdue the effect of huge percentage differences arising through the presence of zero counts. For clarity, for an arbitrary original log-linear parameter estimate q and its corresponding synthetic estimate q^{syn}, the percentage difference was calculated by $100 \times (q^{\text{syn}} - q)/q$

	$\sigma = 0$ (Pois.)	$\sigma = 0.1$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 2$	$\sigma = 5$	$\sigma = 10$
The NBI model							
$\alpha = 0$	−1.7	3.9	−12.0	−0.1	−18.7	10.6	−108.4
$\alpha = 0.005$	−23.5	−30.9	−31.7	−34.8	−34.0	14.8	−32.9
$\alpha = 0.01$	−32.5	−33.2	−33.7	−38.9	−47.8	41.0	−64.7
$\alpha = 0.015$	−38.8	−39.8	−47.7	−52.3	−47.6	−20.3	−3.5
$\alpha = 0.02$	−37.1	−34.0	−44.4	−40.9	−27.7	−42.0	−33.5
The PIG model							
$\alpha = 0$	−1.7	−2.2	16.2	11.2	−33.0	−6.6	187.4
$\alpha = 0.005$	−23.5	−21.7	−28.4	−21.7	−33.0	−73.6	−990.4
$\alpha = 0.01$	−32.5	−29.6	−31.7	−48.6	−42.3	25.9	−399.3
$\alpha = 0.015$	−38.8	−26.4	−36.6	−37.6	−20.6	−64.4	−504.0
$\alpha = 0.02$	−37.1	−47.8	−40.2	−20.6	−40.5	−76.2	425.2

lengths tends to infinity—but the other confidence interval is finite—the overlap value tends to 1/2.

Table 5 presents (trimmed) mean percentage differences between synthetic and observed parameter estimates for various σ and α. Even setting α small can have an adverse effect on utility. For example, when $\sigma = 0$ (the Poisson model), increasing α from 0 to 0.005 causes the (trimmed) mean percentage difference in estimates to fall from −1.7% to −23.5%, thus demonstrating the

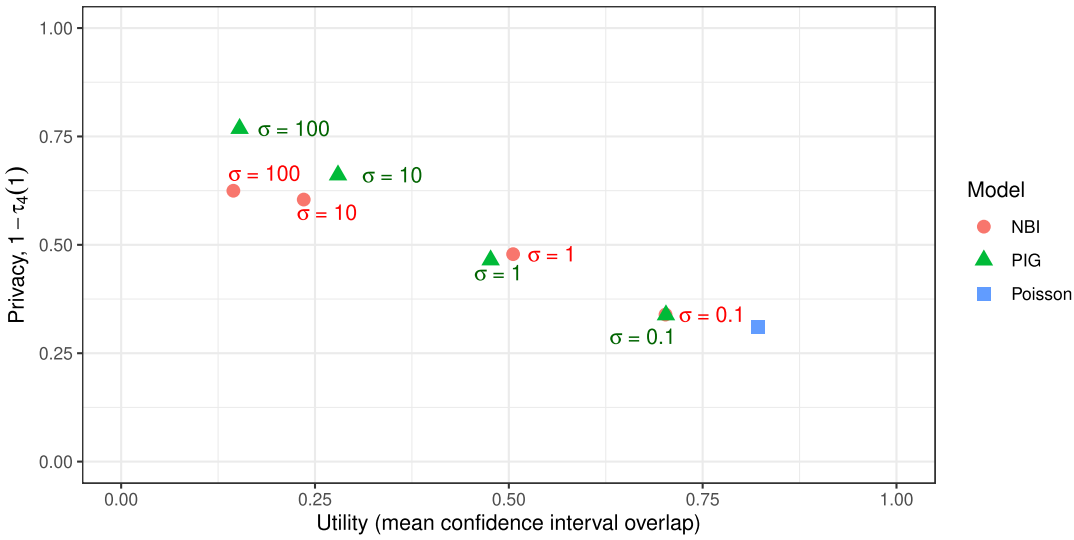

FIGURE 6 This plot, which is resemblant of a product possibility frontier in economics, provides a visual representation of the risk-utility trade-off for different σ ($\alpha = 0$).

bias caused by $\alpha > 0$. The general trend is that increasing α and σ results in larger percentage differences.

4.5 | Balancing risk and utility

A key question a synthesizer would have is: which synthesis method offers the best balance between utility and risk? To address this, the risk-utility trade-off from each generated synthetic data set can be plotted. An example is displayed in Figure 6. Privacy has been measured on the y-axis via $1 - \tau_4(1)$ (that is, $1 -$ risk) and utility on the x-axis by mean confidence interval overlap. The original data sit at the point (1, 0), that is, maximum utility and minimum privacy. All points must lie within the unit square $[0, 1] \times [0, 1]$ and the further from the origin, the better the synthetic data. For instance, when a point is near the origin, it suggests that for the same level of privacy, a greater level of utility is achievable (or vice versa).

This visualisation offers a convenient way to compare the performance of different synthesis models. For example, it may be possible for one synthetic data set to strictly dominate another: the PIG model with $\sigma = 10$ provides greater utility and lower risk than the NBI model with $\sigma = 100$. The choice depends on the priorities of the data holder and users. For example, it may be that synthetic data can only be released if $\tau_4(1)$ is at least 0.5, in which case the synthetic data with the highest utility that satisfies this requirement could be released, here this would be the PIG model with $\sigma = 10$. Alternatively, it may be that only data with a utility value of at least 0.5 would be deemed useful enough for release, in which case the synthetic data generated under a NBI model with $\sigma = 0.1$ would be chosen.

In practice, a range of different metrics for utility and privacy can be created and feed into determining which synthesis method is chosen. This decision is also likely to be application specific.

4.6 | The Poisson model (vs.) the NBI model (vs.) the PIG model

The intention is that the synthesizer would usually use the NBI or PIG distributions, rather than the Poisson, which is far too limited to be used in practice. In general, the NBI and PIG models give similar results, yet this is to be expected as both share the same variance function. Nevertheless, there are some marked differences between the two, especially when σ is large; for example, when $\sigma = 10$ and $\alpha = 0$, there are substantially fewer zeros in the synthetic data when the PIG is used than when the NBI is used ($\tau_1(0)$ values of 0.950 and 0.971 respectively). There is, of course, scope to use other count distributions here; those with a different variance function would have an entirely different profile altogether. Moreover, both the NBI and PIG are also both limited in that they can only model overdispersion and not underdispersion, hence the variance is always greater than the Poisson—and this may be unnecessary. The double Poisson distribution (see Rigby et al., 2019), for example, which can be used to model underdispersed count data, would allow the variance to be set lower than in the Poisson.

5 | DISCUSSION AND FUTURE WORK

In this paper, the case of generating $m = 1$ synthetic data set was considered. But $m > 1$ data sets can be generated using the same framework and, in the same way, certain properties can be found analytically. There might be advantages of doing this, especially in relation to the risk-utility trade-off. It effectively introduces another tuning parameter, thus providing further flexibility. Investigations unreported in this paper have shown promising signs for $m > 1$; for example, particularly for large σ, the gains in utility appear to outweigh the relatively small increase in risk. Moreover, an optimal value for n_{syn}, the sample size of the synthetic data, was not sought. There is scope to set n_{syn} lower or higher than n (the sample size of the original data), which again can be evaluated in relation to the risk-utility trade-off.

While the two-parameter count distributions allow the synthesizer to set the synthetic counts' variance, they cannot control where the variability falls. It is not desirable for the variability to manifest itself in, say, a heavy right tail in the synthesis distribution's probability mass function, because, while some movement is required in synthetic counts to reduce risk, large movements are unnecessary and may have an adverse effect on the data's utility. The use of three-parameter count distributions, such as the Delaporte and Sichel distributions (see Rigby et al., 2019), would provide the synthesizer with control over the skewness in addition to the variance.

This method assumes that the cell counts in the multi-way table are independent. This assumption can be exploited further by specifying different models—for example, different σ and α values—when synthesizing different parts of the original data's multi-way table. Smaller cell sizes could be synthesized using a relatively larger σ than larger cells, which would inject more variability where it is needed.

The method as presented here also assumes that the size of α is constant across all random zeros. However, it can be argued that some random zeros in the original data are more (or less) likely to be non-zeros than others; for example, some zero cells pertain to higher order marginal counts that are also zero. This can be accounted for, to a certain extent, by smoothing the original counts through fitting, for example, an all two-way interaction log-linear model. But the benefits of using saturated models would be lost. The pseudo-Bayes estimator, as presented in

Chapter 12 of Bishop et al. (1975), provides an alternative to adding constant α. A set of prior cell probabilities (denoted by λ) are selected using, for example, external information. Based on these prior probabilities, the observed counts are re-weighted to provide a set of adjusted counts, and a saturated model would then be applied as before, synthesizing from these adjusted values. Hence, while the observed counts are smoothed—and potentially reducing the number of zero cells, thereby helping to minimise the impact of the problem discussed in Section 3.2—they are not smoothed through modelling decisions (setting interactions to zero), but through the choice of λ. This means, however, that the fundamental challenge is just transferred from choosing α to choosing λ. The strategies provided in Bishop et al. (1975) may offer some insights into this, although the objectives of the synthesis would also be relevant here. This is something that would involve further careful consideration and is a substantial research question on its own.

In the empirical study (Section 4), only variables at the pupil level were considered. However, administrative data might have a hierarchical structure that would need to be taken into account. In this example, this could involve incorporating school-level variables into the synthesis. This clearly presents challenges from the modelling and utility perspective, such as ensuring relationships between pupil-level variables within schools are preserved in the synthetic data. However, this also presents interesting questions around disclosure risk because the school-level variables may increase the risk at the pupil-level. The risk and utility challenges associated with multi-level data in this area merit further consideration.

There is no panacea for synthetic data generation. A compromise always needs to be struck between risk, utility and, in the case of large data sets, computational time. Different methods are, of course, suited to different data types and sizes. The conditional approaches outlined in Section 2.1.1, which typically either use GLMs or CART, are effective in synthesizing microdata sets, particularly those comprising a mix of continuous and categorical variables. Yet, when n is large, demands on memory makes it challenging, computationally, to implement such methods. It is more efficient to undertake synthesis of categorical data at the aggregated level. Among these approaches, the advantage of using saturated models is twofold. They eliminate the need to make modelling decisions, which ensures the preservation of relationships, and support an *a priori* approach to synthesis, whereby expected properties of the synthetic data can be established beforehand. This does not just apply to the 'τ metrics' used in this paper, but to many other risk and utility metrics; for example, expressions can be similarly derived—or at least approximated—for general utility measures such as Hellinger distance and Kulback–Leibler divergence. This facilitates a more formal approach to risk and utility that may, in turn, invite greater transparency.

This paper hopefully gives confidence to organisations holding large administrative databases that generating synthetic data is not necessarily a computationally intensive and time-consuming endeavour. Furthermore, the organisations can easily tune the synthesis models in a very transparent way to achieve pre-specified levels of risk and utility.

ACKNOWLEDGEMENTS

This work was carried out as part of a CASE PhD Studentship between Lancaster University and the Office for National Statistics, funded by the Economic and Social Research Council (ESRC). The authors also thank the Associate Editor and the reviewers for their comments and suggestions, which have without doubt improved the article.

DATA AVAILABILITY STATEMENT

The data set used in this paper is openly available in Lancaster University's data repository at http://doi.org/10.17635/lancaster/researchdata/533

ORCID

James Jackson https://orcid.org/0000-0002-4832-6638
Robin Mitra https://orcid.org/0000-0001-9584-8044
Brian Francis https://orcid.org/0000-0001-7926-9085
Iain Dove https://orcid.org/0000-0002-1145-2999

REFERENCES

Agresti, A. (2013) *Categorical data analysis*, 3rd edition. Wiley Series in Probability and Statistics. Hoboken, New Jersey: Wiley-Interscience.

Bates, A.G., Spakulová, I., Dove, I. & Mealor, A. (2019) ONS methodology working paper series number 16 - Synthetic data pilot. Available from: https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/onsmethodologyworkingpaperseriesnumber16syntheticdatapilot [Accessed 10th May 2020].

Bishop, Y.M., Fienberg, S.E. & Holland, P.W. (1975) *Discrete multivariate analysis: theory and practice*. Cambridge: The MIT Press.

Blanchard, S., Jackson, J.E., Mitra, R., Francis, B.J. & Dove, I. (2022) A constructed English School Census substitute. Lancaster University. Available from: https://doi.org/10.17635/lancaster/researchdata/533 [Accessed 17th May 2022].

Breiman, L. (2001) Random forests. *Machine Learning*, 45, 5–32.

Breiman, L., Friedman, J., Stone, C.J. & Olshen, R.A. (1984) *Classification and regression trees*. Boca Raton: Chapman & Hall / CRC.

Caiola, G. & Reiter, J.P. (2010) Random forests for generating partially synthetic, categorical data. *Transactions on Data Privacy*, 3, 27–42.

Deming, W.E. & Stephan, F.F. (1940) On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11, 427–444.

Drechsler, J. (2011) *Synthetic datasets for statistical disclosure control: theory and implementation*, vol. 201. Berlin: Springer (Lecture Notes in Statistics).

Drechsler, J. (2018) Some clarifications regarding fully synthetic data. In: Domingo-Ferrer, J. & Montes, F. (Eds.) *Privacy in statistical databases*. Cham: Springer International Publishing, pp. 109–121.

Drechsler, J. & Reiter, J.P. (2011) An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Computational Statistics & Data Analysis*, 55, 3232–3243.

Duncan, G. & Lambert, D. (1989) The risk of disclosure for microdata. *Journal of Business & Economic Statistics*, 7, 207–217.

Duncan, G.T., Fienberg, S.E., Krishnan, R., Padman, R., Roehrig, S.F. et al. (2001) Disclosure limitation methods and information loss for tabular data. In: Doyle, P., Lane, J., Theeuwes, J. & Zayatz, L. (Eds.) *Disclosure and data access: theory and practical applications for statistical agencies*. Amsterdam: Elsevier, pp. 135–166.

Duncan, G.T., Elliot, M. & Juan Jose Salazar, G. (2011) *Statistical confidentiality principles and practice*. Berlin: Springer Statistics for Social and Behavioral Sciences.

Dunson, D.B. & Xing, C. (2009) Nonparametric Bayes modeling of multivariate categorical data. *Journal of the American Statistical Association*, 104, 1042–1051. Available from: https://doi.org/10.1198/jasa.2009.tm08439

Dwork, C., McSherry, F., Nissim, K. & Smith, A. (2006) Calibrating noise to sensitivity in private data analysis. In: Halevi, S. & Rabin, T. (Eds.) *Theory of cryptography*. Berlin, Heidelberg: Springer, pp. 265–284.

Fienberg, S.E. & Rinaldo, A. (2012) Maximum likelihood estimation in log-linear models. *The Annals of Statistics*, 40, 996–1023.

Goodman, L.A. (1974) Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61, 215–231. Available from: https://doi.org/10.1093/biomet/61.2.215

Graham, P. & Penny, R. (2007) *Multiply imputed synthetic data files*. Official Statistics Research Series (vol. 1). Auckland: Statistics New Zealand.

Hand, D.J. (2018) Statistical challenges of administrative and transaction data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181, 55–605. Available from: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssa.12315

HM Government. (2018) Help shape our future: the 2021 census of population and housing in England and Wales. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/765089/Census2021WhitePaper.pdf [Accessed 10th June 2020].

Hu, J. (2019) Bayesian estimation of attribute and identification disclosure risks in synthetic data. *Transactions in Data Privacy*, 12, 16–19.

Hu, J., Reiter, J.P. & Wang, Q. (2014) Disclosure risk evaluation for fully synthetic categorical data. In: Domingo-Ferrer, J. (Ed.) *Privacy in statistical databases*. Cham: Springer International Publishing, pp. 185–199.

Hu, J., Akande, O. & Wang, Q. (2021) Multiple imputation and synthetic data generation with NPBayesImputeCat. *The R Journal*, 13, 90–110. Available from: https://doi.org/10.32614/RJ-2021-080

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E.S., Spicer, K. et al. (2012) *Statistical disclosure control*. Hoboken: Wiley Series in Survey Methodology.

Information Commissioner's Office. (2020) Guide to the general data protection regulation (GDPR). Available from: https://ico.org.uk/media/for-organisations/guide-to-data-protection-1-1.pdf [Accessed 27th June 2022].

Kaloskampis, I. (2019) Synthetic data for public good. Available from: https://datasciencecampus.ons.gov.uk/projects/synthetic-data-for-public-good/ [Accessed 27th September 2020].

Karr, A.F., Kohnen, C.N., Oganian, A., Reiter, J.P. & Sanil, A.P. (2006) A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician*, 60, 224–232.

Lang, J.B. (1996) On the comparison of multinomial and poisson log-linear models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58, 253–266.

Little, R.J. (1993) Statistical analysis of masked data. *Journal of Official Statistics*, 9, 407–426.

Manrique-Vallier, D. & Hu, J. (2018) Bayesian non-parametric generation of fully synthetic multivariate categorical data in the presence of structural zeros. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181, 635–647. Available from: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssa.12352

Manrique-Vallier, D. & Reiter, J.P. (2014) Bayesian estimation of discrete multivariate latent structure models with structural zeros. *Journal of Computational and Graphical Statistics*, 23, 1061–1079.

Meng, X.-L. (1994) Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 9, 538–558.

Nowok, B., Raab, G.M. & Dibben, C. (2016) Synthpop: bespoke creation of synthetic data in R. *Journal of Statistical Software*, 74, 1–26.

Raab, G.M., Nowok, B. & Dibben, C. (2016) Practical data synthesis for large samples. *Journal of Privacy and Confidentiality*, 7, 67–97.

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

Reiter, J.P. (2003) Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29, 181–188.

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

Reiter, J.P. & Kinney, S.K. (2012) Inferentially valid, partially synthetic data: generating from posterior predictive distributions not necessary. *Journal of Official Statistics*, 28, 583.

Reiter, J.P., Wang, Q. & Zhang, B. (2014) Bayesian estimation of disclosure risks for multiply imputed, synthetic data. *Journal of Privacy and Confidentiality*, 6. Available from: https://journalprivacyconfidentiality.org/index.php/jpc/article/view/635 [Accessed 1st May 2020].

Rigby, R.A., Stasinopoulos, M.D., Heller, G.Z. & De Bastiani, F. (2019) *Distributions for modeling location, scale, and shape: using GAMLSS in R*. Boca Raton: CRC Press.

Rubin, D.B. (1993) Statistical disclosure limitation. *Journal of Official Statistics*, 9, 461–468.

Skinner, C.J. (1992) On identification disclosure and prediction disclosure for microdata. *Statistica Neerlandica*, 46, 21–32.

Skinner, C.J. & Elliot, M.J. (2002) A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64, 855–867. Available from: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9868.00365

Skinner, C. & Shlomo, N. (2008) Assessing identification risk in survey microdata using log-linear models. *Journal of the American Statistical Association*, 103, 989–1001.

Skinner, C., Marsh, C., Openshaw, S. & Wymer, C. (1994) Disclosure control for census micro-data. *Journal of Official Statistics*, 10, 31–51.

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

Stasinopoulos, D.M. & Rigby, R.A. (2007) Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, 23, 1–46.

Taub, J., Elliot, M., Pampaka, M. & Smith, D. (2018) Differential correct attribution probability for synthetic data: an exploration. In: Domingo-Ferrer, J. & Montes, F. (Eds.) *International conference on privacy in statistical databases* vol. 11126, Berlin: Springer (Lecture Notes in Computer Science), pp. 122–137.

Templ, M. (2017) *Statistical disclosure control for microdata methods and applications in R*. Berlin: Springer.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Jackson, J., Mitra, R., Francis, B. & Dove, I. (2022) Using saturated count models for user-friendly synthesis of large confidential administrative databases. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 1–31. Available from: https://doi.org/10.1111/rssa.12876