Using saturated models for data synthesis

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

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

E-mail for correspondence: `j.jackson3@lancaster.ac.uk`

Abstract: The use of synthetic data sets are becoming ever more prevalent, as regulations such as the General Data Protection Regulation (GDPR), which place greater demands on the protection of individuals' personal data, are coupled with the conflicting demand to make more data available to researchers. This paper discusses the approach of synthesizing categorical data at the aggregated (contingency table) level using a saturated count model, which adds noise - and hence protection - to cell counts. The paper also discusses how distributional properties of synthesis models are intrinsic to generating synthetic data with suitable risk and utility profiles.

Keywords: synthetic data; saturated count models; data privacy

1 Introduction

As organisations have both a legal and ethical obligation to protect individuals' personal data, data sets pertaining to individuals cannot be released directly to researchers. Thus prior to release, statistical disclosure control methods, such as the use of synthetic data sets, need to be applied.

Synthetic data sets (Rubin 1993, Little 1993), which are generated by simulating from a model fit to the original data, can be released to researchers in place of the original data. The notion is that, as the synthetic data sets are inherently artificial, individuals' privacy should be protected; while, as synthetic values are based on original values, researchers' ability to obtain valid inferences should remain undiminished. The method relies on the synthesizer – he or she responsible for generating the synthetic data – accurately modelling the data's underlying distribution.

The theory of synthetic data evolved from the multiple imputation of missing data theory (Rubin, 1987). The synthesizer either imputes values for in-

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

dividuals not included in the original data (resulting in fully synthetic data; Raghunathan 2003) or generates replacement values for those individuals who were included in the original data (resulting in partially synthetic data; Reiter 2003). As with imputation, it is typical to release multiple ($m > 1$) data sets to allow analysts – through combining rules; see Drechsler (2011) - to average point estimates and properly account for the extra uncertainty arising from synthesis when calculating estimates' variances.

When synthesizing a data set comprising p variables $Y_1, Y_2, \ldots, Y_p$, the underlying distribution of the data can be captured through a product of conditional models, that is,

$$f(Y_1, Y_2, \ldots, Y_p \mid X) = f_1(Y_1 \mid X) \prod_{j=2}^{p} f_j(Y_j \mid Y_{j-1} \ldots, Y_2, Y_1, X),$$

where X denotes any other data available to the synthesizer, such as other relevant data sets, census tables or administrative data.

The synthesis models for $Y_1, Y_2, \ldots, Y_p$ can take a variety of forms - parametric or non-parametric - ranging from generalised linear models (GLMs), to tree-based methods such as CART, to complex machine learning algorithms. The aim of all these methods, though, is the same: to model the underlying distribution governing the original data.

A categorical data set comprises categorical variables only. Its discrete nature allows the data to be aggregated into a contingency table, such that cell counts give the frequencies with which the various combination of categories (cells) are observed; a given set of categories may not be observed, in which case the cell count would be zero. Synthesis can take place by fitting a *count* model to this table, which is more convenient as the response is univariate rather than multivariate.

2 The motivation for using saturated models

The purpose of synthesis models, then, is for neither inference nor prediction, but to reproduce the structure of the original data. Therefore, unlike when estimating a population parameter, modelling assumptions are not intrinsic to obtaining meaningful estimates and standard errors. For this reason, Jackson et al. (2022) proposed the use of saturated count models for synthesis.

Let $f_1, f_2, \ldots, f_K$ denote the observed counts in the original data's contingency table (the original counts). Then the corresponding counts in the synthetic data's contingency table (the synthetic counts) $f_1^{\text{syn}}, f_2^{\text{syn}}, \ldots, f_K^{\text{syn}}$ are generated by simulating from:

$$f_i^{\text{syn}} \sim X_i \quad i = 1, 2, \ldots, K \tag{1}$$

where X_i is a count distribution with mean f_i. Section 3 considers the best distribution to use for X_i.

The advantage of using a saturated count model is three-fold. Firstly, saturated models require no model selection - which in categorical data involves deciding which interactions to include in the model - as *all* interactions are included. This ensures all relationships are preserved in the resulting synthetic data, thus avoiding the scenario where a researcher's analysis subsequently performed on the synthetic data is more complex than - and hence unsupported by - the synthesis model (Meng, 1994).

Secondly, the time taken to undertake the synthesis, computationally, is substantially reduced because the model-fitting time is null: the model's fitted values are just the original counts.

Thirdly, synthetic counts are unbiased with expectations equal to the original counts. In turn, this gives the synthesizer an insight *a priori* (prior to synthesis) into the likely risk and utility profiles of the synthetic data. To illustrate, original counts of one are usually those at greatest risk of disclosure in a categorical data set because they relate to statistically unique individuals. Therefore, a suitable risk metric for synthetic data is $\tau_3(1)$ (Jackson et al. 2022): the probability that an original count of one is synthesized to one, which relates to a unique in the original data remaining unique in the synthetic data. Now, this unbiasedness property means that if, say, the Poisson is used for synthesis - that is, if the Poisson is chosen as X in (1) - then $\tau_3(1)$ is fixed and equal to exp(-1)=0.37; those familiar with R will recognize this as the quantity *dpois(1,1)*.

This third advantage opens up a new approach in relation to generating synthetic data sets. As synthetic data generation is typically an iterative process, involving extensive post-synthesis evaluations to establish risk and utility, gaining an insight into properties of the synthetic data *a priori* improves the efficiency of the synthesis and invites a more formal approach.

3 The use of multi-parameter count distributions

The most obvious choice of distribution for modelling categorical data is the Poisson. Besides, models often assume that individuals' observations are independent. While for data sets in microdata format this translates into assuming the rows of the data set are independent, for a contingency table it translates into assuming cell counts are Poisson distributed.

The problem with using the Poisson, though, is that each synthetic count's variance is always equal to the mean (the original count). Therefore, the variance of each synthetic count is fixed, and this uncertainty may be insufficient to mask - and hence protect - the underlying original count.

There are benefits, therefore, to using more flexible count distributions instead of the Poisson. The flexibility of the GAMLSS (Generalized Additive Models for Location, Scale and Shape) framework developed by Rigby and Stasinopoulos (2005) is particularly useful here. For more about the distributions mentioned henceforth and their parameterizations, see Rigby et al. (2019), the book written by the creators of the GAMLSS approach.

A two-parameter count distribution such as the negative binomial (NBI) provides the synthesizer with control over the scale (the variance) in addition to the location (the mean), thereby allowing more uncertainty to be applied to original counts. The metric $\tau_3(1)$, for example, then depends on σ the NBI's shape parameter. The intention is that the synthesizer treats σ as a tuning parameter in the synthesis; after all, as the model is saturated, σ could not be estimated anyway through maximum likelihood.

However, increasing the variance of the NBI through increasing σ increases the heaviness of the tails, resulting in a substantial probability point mass at zero. This produces synthetic data with an inflated number of zeros, which is exacerbated by the fact that, as saturated models are used, zero counts in the original data are not synthesized to non-zero counts.

This calls for further flexibility and motivates the use of three-parameter count distributions, which allow the synthesizer to control the shape in addition to the location and scale. One such example is the Delaporte distribution. For a given mean and variance, the shape of the Delaporte can be adjusted to reduce the heaviness of the tails, resulting in fewer zero synthetic counts as well as fewer unnecessarily large synthetic counts. This can be seen in Figure 1, which gives three Delaporte distributions, with the same means and variances but different shapes; for example, the probability of obtaining a zero is much greater in the distribution given by the red (solid) line than in the other two.

The problem in general, though, with distributions that arise through Poisson mixtures (such as the NBI and Delaporte), is that their variances are *increasing* functions of the mean, hence relatively more noise is applied to larger counts than smaller counts. However, as larger counts tend to be lower risk than smaller counts, it is preferable if the variance is a *decreasing* function of the mean, so that larger counts are perturbed less.

Rather than using a standard count distribution, an alternative it to use discretization to produce a more bespoke count distribution, by discretizing a continuous distribution defined on the interval $(0, \infty)$ - an "underused" method according to Rigby et al. (2019). A candidate for discretization is the gamma family (GAF) distribution, which has three-parameters μ, σ and ν, and where ν controls the variance-mean relationship. The mean is μ and the variance $\sigma^2\mu^\nu$; thus, when $\nu < 0$, the variance is a decreasing function of the mean, and larger counts *are* perturbed less than smaller counts - the desired behaviour. Figure 2 displays the variance-mean relationship for three GAF distributions, which is one of exponential decay, where ν controls the rate at which the variance falls away.

4 Conclusion

To briefly conclude, while saturated models are uninformative from an inferential perspective and too rigid from a predictive perspective, they have

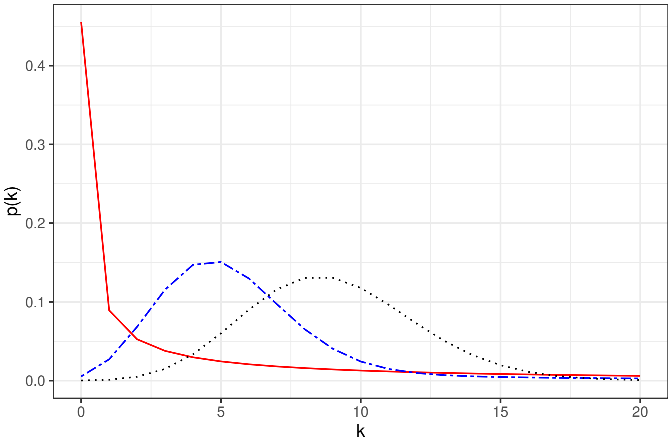

FIGURE 1. The probability mass functions of three Delaporte distributions with the same mean and variance, 10 and 510, respectively. The flexibility afforded by a three-parameter count distribution allows the shape of the distribution to be adjusted. Incidentally, the red (solid) line is an NBI distribution, which is a special case of the Delaporte.

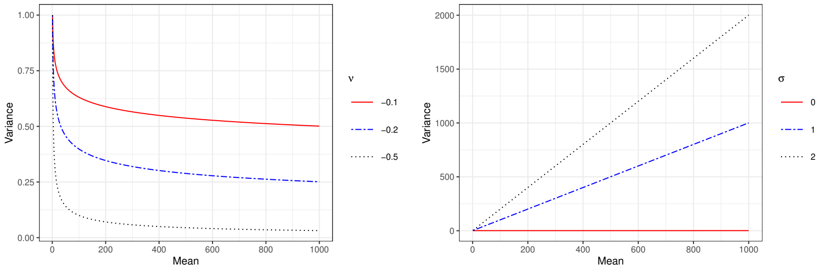

FIGURE 2. The variance-mean relationship for three GAF distributions with different ν values (with $\sigma = 1$). For comparison, the variance-mean relationship for three NBI distributions are placed alongside (with different σ values).

a practical use in data synthesis, where it suffices to obtain a noisy version of the original data. Coupled with the use of a flexible multi-parameter count distribution - for which it can be equally difficult to justify the use of in practice - saturated models allow properties of the synthetic data to be derived analytically *a priori*, thus facilitating a more efficient and transparent synthesis.

References

Drechsler, J. (2011). *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation.* Lecture Notes in Statistics. New York: Springer.

Jackson, J. E., Mitra, R., Francis, B. J., Dove, I. (2022). Using saturated count models for user-friendly synthesis of categorical data. *Forthcoming in Journal of the Royal Statistical Society: Series A (Statistics in Society).* Preprint available at https://arxiv.org/abs/2107.08062

Little, R. J. (1993). Statistical Analysis of Masked Data. *Journal of Official Statistics*, **9**, 407 – 426.

Meng, X.-L. (1994). Multiple-Imputation Inferences with Uncongenial Sources of Input. *Statistical Science*, **9**, 538 – 558.

Raghunathan, T.E., Reiter, J.P., Rubin, D.B. (2003). Multiple Imputation for Statistical Disclosure Limitation *Journal of Official Statistics*, **19(1)**, 1 – 16.

Reiter, J.P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology*, **29(2)**, 181 – 188.

Rigby, R. A., Stasinopoulos, M. D. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **54(3)**, 507 – 554.

Rigby, R. A., Stasinopoulos, M. D., Heller, G. Z. and De Bastiani, F. (2019) *Distributions for Modeling Location, Scale, and Shape: Using GAMLSS in R.* CRC Press.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley.

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