Statistical Journal of the IAOS 32 (2016) 267–269
DOI 10.3233/SJI-160999
IOS Press

Comments on four papers on synthetic data in Volume 32 Issue 1 the Statistical Journal of the IAOS

Gillian M. Raab
Administrative Data Research Centre – Scotland, University of Edinburgh, UK
E-mail: Gillian.raab@ed.ac.uk

One of several explanations of why Homo Sapiens is the only surviving sub-species of the genus Homo is the extended length of our childhood and adolescence. The value of this extended maturation and developing period may be that it allows us to learn and carry out complex tasks. Like Homo Sapiens, methodology for synthetic data has had a long learning period. The idea of using synthetic data for disclosure control was conceived more than 20 years ago [1–3], but it was a further 10 years before the first papers describing how to do it appeared in the literature [4,5]. The subsequent decade was one of rapid development and innovation when the methodology was tested and expanded. The energy and enthusiasm for synthetic data of Reiter and his colleagues was responsible for many major developments; see the monograph by Drechsler [6] for a review. Towards the end of synthetic data's second decade real applications began to appear [7–9]. Two of the four substantial papers that deal with synthetic data in this issue [10,11] are examples of mature methodology, while the other two [12,13] deal with disclosure control, the aspect of synthetic data that is at an early stage in its development. My comments here are from the point of view of a practitioner looking for useful and workable ideas in this field. Our project to provide data for the UK Longitudinal Studies (LSs) is referred to in Vilhuber et. al.'s overview of international developments [14]. More details of our methods and our *synthpop* package for R are available [15–17].

The paper by Vilhuber and Miranda [10] is a particularly good example of a new method developed in response to a real need. It is rare for a journal article to illustrate the difficulties that arise when statistical procedures are used on real data with all their imperfections and idiosyncrasies. The idea in this paper is to produce tabulations from synthetic data and use them to replace cells in tables from the original data that have been suppressed because of perceived disclosure risk. Suppressed cells in the open-source tabulations, Business Dynamics Statistics[1] (BDS), are replaced with those from tabulations of the synthetic Longitudinal Business Data (synLBD).[2] The authors take advantage of the fact that the BDS and the synLBD are both derived from the same original data sources. Their tentative conclusions are somewhat disappointing, showing that using the synLBD does not improve validity compared to a much simpler method of adding multiplicative noise to the cells. But the reasons for this relate to known problems with the utility of the synLBD for the very statistics that the BDS tabulates, and also to differences between the versions of the BDS and synLBD that were available to the authors. The problems with the synLBD have been remedied by a new methodology [18] that will be used for future versions. Their methods may work better when the synthetic data are produced specifically for the purpose of replacing cells in tables. In particular, the methods based on their algorithm 2, which overcome problems of inconsistent

[1] See https://www.census.gov/ces/dataproducts/bds/data.html.
[2] See https://www.census.gov/ces/dataproducts/synlbd/.

margins, may prove to be the best solution. We await the new synLBD to see if this is the case.

The paper by Wei and Reiter [12] is also concerned with preserving marginal totals. They use a mixture of Poisson distributions as their modelling technique. They preserve the margins by utilising the fact that each of the components of their mixture becomes a multinomial distribution when conditioned on the sum of the components. Since the complete model is the sum of the components of the mixture, the overall marginal totals are also preserved. Mixture models have increased in popularity in recent years for their flexibility in modelling awkward distributions [19]. Wei and Reiter show that this is the case for their example, the synthesis of data on numbers of employees and their wages from the Colombian business survey. They carry out a very thorough investigation of the disclosure risk for these data, focussing on the possibility of identifying extreme values. Modelling the data with either a Poisson or Multinomial model resulted in unacceptably high probabilities of disclosure for the extreme values in this data set. Unfortunately their modification of the methods to avoid this, by collapsing the upper tails of the distributions, damages the utility of the data. But these conclusions are perhaps too dependent on the properties of a single small ($n = 1051$) data set with extremely skewed variables. The extreme values are not only problematic for disclosure but are highly influential for statistics such as correlations and regression coefficients; in fact one might question the use of such parametric methods for these data. There could be other examples where one might wish to preserve marginal totals where their methods might work better. An example in our own recent work has been the need to synthesise the totals of different types of individuals in a household (numbers of resident adults and children, servants, visitors, etc.) while preserving the total household size. In their example Wei and Reiter synthesise only one set of Multinomial variables at a time, while most synthetic data applications use a sequence of conditional models. Although the Poisson or Multinomial mixture distributions can be extended to include covariates , this might compromise the ability of the Multinomial model to preserve the marginal totals, depending on how they were incorporated into the sequence of synthesising models. In conclusion, these authors have presented what appears to be a very promising addition to our toolkit for data synthesis but have evaluated it on a particularly problematic example.

When synthetic data were first proposed it was assumed that there would be little or no disclosure risk if all the data were replaced by synthetic values. That this is not the case is illustrated by the analysis of the Colombian data discussed above. Where specific examples of synthetic data have been evaluated for disclosure control the results have been encouraging, but no general measures that work for synthetic data have been derived. Differentially private data release mechanisms are usually produced by adding a controlled random noise to tabulations, where the amount of noise is inversely related to the ε parameter that controls privacy. A small ε will protect privacy, but add a large amount of noise and vice-versa. This model does not accord well with the usual method of producing synthetic data where the noise is determined by the residuals from the model used to generate the synthetic data. Previous attempts to produce differentially-private synthetic data [20–22] have controlled ε by imposing an informative prior on the generating distribution. Unsurprisingly, this results in reduced utility for the synthetic data. The paper by Schmutte [13] generates synthetic data from differentially private tabulations by a complex iterative scheme that is very different from the usual methods used to produce synthetic data. While this method produces reasonable utility for the tabulations used in the data generation, it performs poorly for other tabulations that are not defined by them. This is not news to those of us who produce synthetic data, since we know that they are only as good as the models used to generate them. Yet this method, if it can be extended in some way, may offer a means of controlling disclosure for synthetic data.

The paper by Maclure and Reiter [12] offers what may be a more immediate way forward for those of us producing synthetic data. Using assumptions about an intruder's knowledge of the original data, similar to those used in differential privacy, they compute the probability of disclosure for individual observations. Even for the two simplified examples they present their methods are computationally intensive. But the goal of identifying risky observations is a good one and a way of adapting these methods for larger and more complex data sets would be very worthwhile.

Overall these four papers have much to offer those of us on the practical side of producing synthetic data. Some of it is immediately useful, but we may have to wait while other techniques develop further before they can deliver contributions to practice.

Funding

This work was supported by the Economic and Social Research Council grant number ES/L007487/1 (Administrative Data Research Centre – Scotland).

References

[1] D.B. Rubin, Discussion of statistical disclosure limitation, *Journal of Official Statistics* **9**(2) (1993), 461–468.
[2] R.J. Little, Statistical analysis of masked data, *Journal of Official Statistics* **9**(2) (1993), 407–426.
[3] S.E Fienberg, A radical proposal for the provision of micro-data samples and the preservation of confidentiality, Technical report, Department of Statistics, Carnegie-Mellon University. (1994).
[4] J.P. Reiter, Satisfying disclosure restrictions with synthetic data sets, *Journal of Official Statistics* **18**(4) (2002), 1–19.
[5] T.E. Raghunathan, J.P. Reiter and D.B. Rubin, Multiple imputation for statistical disclosure limitation, *Journal of Official Statistics* **19**(1) (2003), 1–16.
[6] J. Drechsler, Synthetic Datasets for Statistical Disclosure Control Theory and Implementation, New York: Springer, 2011.
[7] J.M. Abowd, M. Stinson and G. Benedetto, Final Report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project. U.S. Census Bureau; (2006). Available from: http://www2.vrdc.cornell.edu/news/?p=308.
[8] S.K. Kinney, J.P. Reiter, A.P. Reznek, J. Miranda, R.S. Jarmin and J.M. Abowd, Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database, *International Statistical Review* **79**(3) (2011), 362–384. Available from: http://ideas.repec.org/a/bla/istatr/v79y2011i3p362-384.html.
[9] J. Drechsler and L. Vilhuber, A first step towards a German SynLBD: Consructing a German Longitudinal Business Database. *Statistical Journal of the IAOS* **30**(2) (2014), 137–142.
[10] J. Miranda and L. Vilhuber, Using partially synthetic microdata to protect sensitive cells in business statistics, *Statistical Journal of the IAOS* **32**(1) (2016), 69–80.
[11] L. Wei and J.P. Reiter, Releasing synthetic magnitude micro-data constrained to fixed marginal totals, *Statistical Journal of the IAOS* **32**(1) (2016), 93–108.
[12] D. MacLure and J.P. Reiter, Assessing disclosure risks for synthetic data with arbitrary intruder knowledge, *Statistical Journal of the IAOS* **32**(1) (2016), 109–126.
[13] I.M. Schmutte, Differentially private publication of data on wages and job mobility, *Statistical Journal of the IAOS* **32**(1) (2016), 81–92.
[14] L. Vilhuber, J.M. Abowd and J.P. Reiter, synthetic establishment data around the world, *Statistical Journal of the IAOS* **32**(1) (2016), 65–68.
[15] B. Nowok, G.M. Raab and C. Dibben, synthpop: Bespoke creation of synthetic data in R, *Journal of Statistical Software*. Forthcoming. (2015). Available from https://cran.r-project.org/web/packages/synthpop/vignettes/synthpop.pdf.
[16] B. Nowok, G.M. Raab and C. Dibben, Assisted methods for providing bespoke synthetic data for the UK longitudinal studies and other sensitive data, *Statistical Journal of the IAOS*. Submitted (2016).
[17] G.M. Raab, B. Nowok and C. Dibben, Practical synthesis for large samples. Submitted (2016). Available from http://arxiv.org/abs/1409.0217.
[18] S.K. Kinney, J.P. Reiter and J. Miranda, SynLBD 2.0: Improving the synthetic Longitudinal Business Database, *Statistical Journal of the IAOS* **30**(2) (2014), 129–135.
[19] G. McLachlan and D. Peel, Finite Mixture Models, Wiley, New York, 2000.
[20] J.M. Abowd and L. Vilhuber, How protective are synthetic data? in: *Privacy in Statistical Databases*, J. Domingo-Ferrer and Y. Saygun, eds, New York: Springer-Verlag, 2008, pp. 239–246.
[21] A.S. Charest, How can we analyze differentially-private synthetic datasets, *Journal of Privacy and Confidentiality* **2**(2) (2010).
[22] D. McClure and J.P. Reiter, Differential privacy and statistical disclosure risk measures: an investigation with binary synthetic data, *Transactions on Data Privacy* **5**(3) (2012), 535–552.