Journal of Official Statistics, Vol. 24, No. 2, 2008, pp. 319–321

Letter to the Editor

Letters to the Editor will be confined to discussion of articles which have appeared in the Journal of Official Statistics.

Dear Readers of *JOS*,

I write to comment on the article by Muralidhar and Sarathy (MS) that appeared in the September 2006 issue of JOS. MS compared a particular form of data perturbation to multiple imputation as methods for limiting the risks of identification disclosures in public use data. Their primary conclusion was that the perturbation techniques are more effective than the multiple imputation techniques because they exactly preserve means and covariances, whereas the multiple imputation approach preserves these quantities up to random noise.

While I compliment MS for comparing different methods of disclosure limitation and for writing a clear and interesting paper, I believe their evaluation was too limited to adequately compare the two approaches. The comparison does not justify the recommendation that the particular method of data perturbation should be preferred to multiple imputation, as I describe below.

Multiple imputation is designed to handle categorical, continuous, or mixed data. For example, research has shown inference valid simulation of categorical identifiers like race, sex, and marital status using various forms of logistic regression (Reiter 2005a). It also has been used to protect elaborate, relationally-linked data products where specification of the complete data-generating process is not feasible (Abowd and Woodcock 2002). In contrast, the perturbative techniques of MS operate only on continuous variables. They are not appropriate for datasets with many categorical and mixed variables.

Imputation models can mimic the distributions of the data; they need not be confined to linear regressions or normally distributed errors. For example, recent synthesis projects for mixed and highly skewed data are based on CART models (Reiter 2005b) and density regressions (Abowd and Woodcock 2004). These projects show that it is possible (with reasonable mean squared error) to preserve univariate distributions, maintain interaction and nonlinear effects, and enable valid estimation of sub-domain relationships. In contrast, MS do not provide evidence that the perturbative methods can preserve these and other fine features of distributions.

Even preserving means and covariances may not guarantee that the analyst obtains the same results from the released and original data. If the distributional features are badly distorted with perturbed data, the analyst of the perturbed data could arrive at a model that fits poorly on the original data, because the model diagnostics based on perturbed data may suggest entirely different (and inappropriate) models. This issue has received little attention in the evaluations of disclosure limitation methods.

The multiple imputation approach is, or at least can be, relatively transparent to the public analyst. The agency can release meta-data describing the imputation models. When the analyst seeks to estimate relationships that are not included in the imputation models,

for example certain higher order interactions, the analyst knows from the meta-data whether or not she can use the multiply-imputed data. In contrast, it is very difficult for the analyst of noise-perturbed data to determine how much a particular analysis has been distorted (unless it depends only on the first two moments of the continuous variables).

Multiple imputation – for disclosure limitation or for missing data – is a general purpose tool. For any one purpose, such as exactly preserving means and covariances, other tools such as the data perturbation of MS will outperform multiple imputation. However, when considering the wide variety of analyses done with public use data, a general purpose tool that has the possibility of preserving many different types of relationships is arguably the best we can do.

After all, if we are worried only about analyses based on the first two moments of the entire dataset, why not just release the mean vector and covariance matrix without any microdata?

Sincerely,

Prof. Jerry Reiter
Institute of Statistics and Decision Sciences
Duke University
Durham, NC 27708
U.S.A.
Email: *jerry@stat.duke.edu*

References

Abowd, J.M. and Woodcock, S. (2002). Disclosure Limitation in Longitudinal Linked Data. In Confidentiality, Disclosure, and Data Access: Theory and Practial Applications for Statistical Agencies, P. Doyle, J. Lane, L. Zayatz, and J. Theeuwes (eds). Amsterdam: North Holland, 215–277.

Abowd, J.M. and Woodcock, S. (2004). Multiply-Imputing Confidential Characteristics and File Links in Longitudinal Linked Data. In Privacy in Statistical Databases PSD2004, J. Domingo-Ferrer and V. Torra (eds). Berlin: Springer-Verlag, 290–297.

Reiter, J.P. (2005a). Releasing Multiply-imputed, Synthetic Public Use Microdata: An Illustration and Empirical Study. Journal of the Royal Statistical Society, Series A, 168, 185–205.

Reiter, J.P. (2005b). Using CART to Generate Partially Synthetic Public Use Microdata. Journal of Official Statistics, 21, 441–462.

Reply

Dear Readers of *JOS*,

This is our response to Professor Reiter's comment on our article titled "A Comparison of Multiple Imputation and Data Perturbation for Masking Numerical Variables" that appeared in the September 2006 issue of JOS.

In our opinion, the bulk of Professor Reiter's letter has only marginal relevance to the article we published, since it deals with many situations other than what we have addressed in our article. As should be evident from the title of our article, we are dealing with protecting confidential numerical variables. In that context, we have irrefutably shown that the sufficiency-based perturbation (SBP) approach offers a better alternative to multiple imputation (MI) when the goal is to maintain important aggregate statistics like the mean vector and covariance matrix. There are certainly other situations, as pointed out by Professor Reiter, where MI and its variations may be applicable. It is also true that there are other methods such as Data Shuffling (Muralidhar and Sarathy 2006) that are applicable in those situations, such as in maintaining nonlinear relationships and marginal distributions. Hence, those issues should be debated outside the scope and context of the current article. Professor Reiter's letter *offers no reason why SBP should not be preferred over MI* in the context addressed in our article.

The one specific criticism was his comment that "if we are worried only about analyses based on the first two moments of the entire dataset, why not just release the mean vector and covariance matrix without any microdata?" This criticism is surprising since our demonstration of the superiority of SBP over MI in maintaining these statistics were based on the same examples used in Professor Reiter's research to illustrate the efficacy of MI. What the results of our article in JOS show is the following. For the very same cases used to illustrate the efficacy of MI, releasing the SBP microdata is at least as effective as releasing just the mean vector and covariance matrix. In MI, even after analyzing multiple data sets (perhaps as many as 100), the results of the analysis yield *less information with more effort* than is available from the mean vector and covariance matrix. This is precisely why we should prefer SBP over MI.

Reference

Muralidhar, K. and Sarathy, R. (2006). Data Shuffling: A New Approach for Masking Numerical Data. Management Science, 52, 658–670.

Prof. Krishnamurty Muralidhar
University of Kentucky
Gatton College of Business & Economics
425 G Business & Economics Bldg
Lexington, KY 40506-0034
U.S.A.
Email: krishm@uky.edu

and

Dr. Rathindra Sarathy
Oklahoma State University
Spears School of Business
Stillwater, OK 74078
U.S.A.
Email: sarathy@okstate.edu