DISCUSSION

Nancy Spruill, Office of the Assistant Secretary of Defense
for Force Management and Personnel

I enjoyed both papers. They both contain some good ideas for providing confidentiality for individual data.

KIM PAPER

Protection of Taxpayer Confidentiality with Respect to the Tax Model

Jay's paper proposes a modification to the traditional masking technique of adding random noise. He is interested in how his transformed data performs in regression analyses.

The noise he proposes adding has zero means and has variance-covariance structure proportional to that of the underlying data.

The advantage to his procedure is that the first and second moments of the transformed data are the same as those of the original data.

But what about the regression analyses, one of the most common techniques that uses these data? In the usual case when the underlying variances are unknown, the means of the regression coefficients are the same as they would have been if we had done the regression on the original data. However, although Jay gives the sampling variance for the regression data, he does not predict the variance of the regression coefficients compared to the variance of the coefficients using the original data. This might be difficult analytically, but showing results for some randomly generated samples might be useful to the data user. I would suggest adding this to the paper.

Jay does use some random samples to compare his technique with another for adding random noise. This other technique adds random noise for each observation that has mean zero and variance equal to 1/2 the variance of that observation, not 1/2 the variance of the population. He shows how his technique for adding random noise performs. It appears superior to the other technique. However, Jay might want also to compare it to the more common way of adding random noise -- the same as his technique, except no correlation among the error variables.

Jay's technique shows real promise. As Jay mentioned, the next step is to try to "reidentify" the data. Also, he needs to address two issues:

First, how do you handle zero values? Do you want to add random noise to them? If so, how does this affect researchers? If not, can the pattern of zero and non-zero values be used to identify individuals?

And second, how does this technique handle outliers? Are they getting enough protection? If not, can we combine this technique of adding random noise as Jay proposed with some other technique or can he modify this technique to give them protection too?

STRUDLER-OH-SCHEUREN PAPER

A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation

The IRS Individual Tax Model is an important asset for researchers. Also, Mike's paper is especially important in light of the renewed interest in the taxpayer's behavior under the new tax laws. Researchers and special interest groups will want to see the effects of changing the tax rules.

The paper addresses an important issue -- can several data elements taken together increase the chances of identifying specific individuals in the Tax Model? His paper proposes several techniques for further protecting the confidentiality of taxpayers who are sampled for the Tax Model. For completeness, the paper should include an estimate of the number of potential disclosure problems without any improvements. Mike defined a disclosure problem as a cell containing 2 or less individuals. He proposes several changes to increase protection.

First, he eliminated certain types of readily accessible data, such as alimony paid and blindness codes; and, second, he altered other types of easily accessible data, such as age exemptions and number of children.

These changes reduced the amount of data with potential problems to 5.7% -- all in the highest income group.

Next, Mike further modified the data by reducing the upper income groups from 100% sampling to 33% sampling and divided the high income sampled taxpayers into 35 categories. Within these categories, he deleted some outliers; then, he blurred, or grouped, the data -- 3 at a time. Mike found no disclosure problems after using these additional techniques. However, his results table (#2) contains only results for one of his 11 subsamples. His technique also introduced bias. He states that the bias is predictable and the user can possibly adjust his statistics accordingly.

In his paper, when Mike is looking for disclosure problems, he is looking for cells with 1 or 2 entries. He dismisses attempts to link back data for individuals, as I proposed in my earlier work based on IRS' research showing that data were rarely known exactly [1]. I don't think this means he shouldn't try to "reidentify" the data of individuals. I think he should modify the technique and use rounded data or some other kind of modified data as the "true value." But he should still look for individual matches. This seems like a reasonable adjustment and, although Mike's research gives us a high degree of confidence in the data protection, this additional look would allow researchers and individual taxpayers to better address the question of whether an individual can be identified.

FOOTNOTE

[1] Spruill, Nancy. (1983) "The Confidentiality and Analytic Usefulness of Masked Business Microdata," 1983 American Statistical Association Proceedings, Section on Survey Research Methods, pp. 602-607.