

This is a working document prepared by the Energy Information Administration (EIA) in order to solicit advice and comment on statistical matter from the American Statistical Association Committee on Energy Statistics. This topic will be discussed at EIA's spring 2006, meeting with the Committee to be held April 6 and 7, 2006.

Measuring Perceptions of Applying Alternative Disclosure Limitation Methods
Jacob Bournazian
SMG, EIA
jacob.bournazian@eia.doe.gov

OVERVIEW

Suppression is the most common method that federal agencies use to protect the confidentiality of reported data when releasing an information product. During the past 15 years, alternative disclosure limitation methodologies have been developed for protecting tabular and microdata. These methodologies offer new options in releasing data products for statistical agencies to protect the confidentiality of the reported data. These alternative methods offer different levels of protection of sensitive information. They also impact significantly on the utility of the information provided to users because each method uses a different approach for modifying the reported data. Research is needed to measure the perceptions of the data user community and the survey respondents of applying alternative disclosure limitation methods to confidential Energy Information Administration (EIA) data.

Before implementing any suppression procedures it would be useful for EIA to understand the impact that these methods would have on our users and survey respondents. Particularly it is important to understand how our data users would perceive the usefulness of data to meet their needs under different suppression scenarios. Likewise, it is also important to understand how survey respondents perceive the use of alternative disclosure limitation methods and whether the use of these methods will affect the quality or accuracy of survey responses.

ISSUES

- I. How to design a study to measure the perceptions of data users on the data utility of information products that are protected by disclosure limitation methods other than cell suppression?
- II. How to design a study to measure the perceptions of survey respondents on their willingness to report accurately when they are informed that information products are protected by disclosure limitation methods other than cell suppression?
- III. What types of users should be included in the study? Are there important classes or types of users and survey respondents that need to be included?
- IV. What mode(s) of collecting feedback should be used in the study?

MEASURING BUSINESS PERCEPTIONS

For the past 20 years, the Census Bureau has measured the perceptions of household respondents on privacy, confidentiality, and data sharing. The types of research studies that measured household perceptions included focus groups, individual cognitive research, and mail surveys to groups of individuals. (Mayer, 2000). The studies of the household respondents over the past 20 years suggest that the public doubts the confidentiality pledge the agency provides and there is an underlying tendency to believe that the information is shared with other federal agencies. In contrast, a large body of literature relating to the perceptions of business respondents does not exist and very little research has been done to specifically measure the perceptions of businesses relating to data confidentiality.

One of the first studies on business perceptions of data confidentiality was through the Survey of Business Perceptions of Confidentiality that was sponsored by the U.S. Census Bureau and conducted by the Urban Institute. (Greenia et. al., 2001). This study focused on two issues:

- 1) What kinds of information do businesses consider sensitive – and for how long are they perceived to be sensitive?
- 2) What are business perceptions of an agency's ability to collect and protect data?

A sample of 5,000 companies was drawn from a frame of 11.3 million businesses that was obtained from Dun & Bradstreet. The sample was stratified into 4 strata with 1,250 businesses in each stratum. The cutoffs for the strata were based on the number of employees: 0-49 employees; 50-249 employees, 250-499; and 500 plus. Of the 5,000 companies, 2,530 were multi-unit business with headquarter locations and 2,470 were single location businesses with only one business. The target respondent was an authority figure in the business.

There were only 509 useable responses from this survey resulting in an overall response rate of 10.6%. Response rates declined as business size increased. Multiple factors may have contributed to the low response rate. The survey was voluntary, conducted by the non-federal agency contractor, and it was conducted during the holiday season. Although response rates varied according to size of the firm, there was no significant relationship with size or industry for most of the survey responses.

The results of the survey of Business Perceptions of Confidentiality showed that the level and duration of sensitivity varied across different data elements on a survey. Business respondents were similar to household respondents with an even split in their beliefs that federal agencies kept their reported values confidential. The study also showed that the more confidence a respondent had concerning government competence, the less concern they had about providing data to either statistical or regulatory agencies, and likewise, as trust in the federal agency increased, the respondent's concern over reporting information decreased.

ALTERNATIVE METHODS TO PROTECT TABULAR DATA

There are alternative disclosure limitation methodologies that may be used to protect tabular data. Noise addition, data swapping, and controlled tabular adjustment are three methods that are discussed in this paper. For each of these three methods, confidential data are protected by modifying the data in some manner. Noise addition and data swapping modify the microdata prior to tabulation. Controlled tabular adjustment modifies the aggregate data after the cell values have been tabulated. Noise addition and data swapping are two methods that have been implemented by a federal statistical agency.

Other methods are also available for protecting tabular data such as publishing a range rather than suppressing the table cell value. Collapsing over table categories is another method that may be applied to avoid suppressing sensitive values while providing the user with some useful information. These two methods do not disturb the data and therefore do not affect the relationships between variables. However, like cell suppression, there is some information loss to the table.

Although each of the three methods discussed below allow more data to be published, each of these methods disturbs the relationships between variables in a table. This could affect regression results by affecting variable coefficients and their significance in a model.

1) ADDING NOISE TO MICRODATA PRIOR TO TABULATING DATA - Adding noise to the underlying microdata is a method that has been used by the U.S. Census Bureau on their Research and Development Survey and by the National Agricultural Statistics Service to data released from the Agricultural Resource Management Survey. It is different from the noise procedures used to protect public-use microdata files. This noise addition method adjusts each value by a small amount (the exact percent to remain confidential within the statistical agency) prior to tabulating the aggregate table cell values. Each company reporting in the sample or survey is assigned a multiplier or noise factor. A company may have several different stores. In this case, each store may be assigned a slightly different multiplier as long as the overall distribution of the multipliers across all stores within a company average the specified percent for adjusting that company's reported values. Noise addition relies on the random assignment of the multiplier to control the effects of adding noise to different types of cells.

For example, if a company's data is adjusted by 10%, then its' data would be multiplied by a number that is close to either 1.1 or 0.9. Any type of distribution can be used to choose the multipliers for each store. In this example, whatever distribution is used to generate a multiplier of 1.1, it is important that the same distribution shape, or its "mirror image," be used to generate the multipliers near 0.9 to adjust data in the opposite direction. The two distributions of multipliers should produce a joint distribution of multipliers that is symmetrical and approximates 1.

The direction of adding the noise to each responding company is **randomly assigned** so that one company's data may be adjusted upward by 10% and another company's data may be adjusted downward by 10%. Using the example of 10% as the base for perturbation, this is equivalent to determining if all stores in a company have multipliers close to 1.1 or close to 0.9. The next step in the process is to **randomly assign** a multiplier to each store within a company. The multipliers would be generated from that half of the overall distribution of the multipliers that corresponds to the direction of perturbation assigned to that company.

An example of assigning multipliers to a set of respondents is as follows:

<u>Company</u>	<u>Store</u>	<u>Direction</u>	<u>Multiplier</u>
Company A		1.1	
	Store A1		1.12
	Store A2		1.09
	Store A3		1.10
	Store A4		1.11
Company B		0.9	
	Store B1		0.89
	Store B2		0.93
Company C		1.1	
	Store C1		1.08

In this example, the expected value of the amount of noise added in any cell value is zero because of the symmetry of the distribution of the multipliers and the **random assignment** of both the direction of perturbation and the multipliers within each company. The probability that a company's stores will be perturbed in a positive direction is equal to the probability that they will be perturbed in a negative direction. The distribution of the multipliers is symmetric about 1. The expected value of any given multiplier is 1, hence the expected value of the *amount* of noise in any given store is 0, and the amount of noise in any cell value is simply the sum of the noise in its component stores.

2) DATA SWAPPING – Data swapping was used by the U.S. Census Bureau for some data products generated from the Census 2000. The procedure was performed on the underlying microdata, and all tabulations from the 100% (short form) and from the sample (long form) data were created from the swapped files. It affected pairs of households (or partnered households) where one or both of those households had a high risk of disclosure. The selection process can target those records with the most disclosure risk. The set of census households that were deemed as having a disclosure risk were selected from the internal census data files. These households were unique in their geographic area based on certain characteristics. The data from these households were swapped with data from partnered households that had identical characteristics on a certain set of key variables but were from different geographic locations. The households that were swapped were kept confidential. The swapping procedure was performed

independently for the 100% block data and the sample data. The swapping rate can vary across regions. To maintain data quality, there was a maximum percent of records that were swapped for each state for the 100% data and another maximum percent for the sample data. All tabular data products are then created from the swapped file.

3) CONTROLLED TABULAR ADJUSTMENT - Controlled tabular adjustment is a relatively new approach, similar to controlled rounding, that is applied to tables of magnitude data. With controlled tabular adjustment, each original sensitive value of a table is replaced with a safe value that is a “sufficient distance” away from the true value; and non-sensitive cell values are minimally adjusted to ensure that the published marginal totals are additive. A “sufficient distance” from the true value would be the value that needs to be added to the cell total so that the cell value is no longer sensitive. Less adjustment is needed to the internal cells of the table if the marginal is also adjusted. This method has not been implemented by any federal agency. The table below illustrates how this method is applied to protect tabular data.

Example – A Table Protected by Controlled Tabular Adjustment

Sales of No. 2 Distillate Fuel (Million gallons/day)

Region	Residential	Commercial	Industrial	Agricultural	Total Retail Sales
East	15	5 - 2 = 3	3	1* + 1 = 2	24 - 1 = 23
West	10	10	10	15	45
South	3	10	10	2 - 2 = 0	25 - 2 = 23
North	12	14	7	2 + 1 = 3	35 + 1 = 36
U.S.	40	39 - 2 = 37	30	20	129 - 2 = 127

Controlled tabular adjustments to individual cell values are shown in **Bold** font.

SUMMARY

There are alternative disclosure limitation methods, besides cell suppression, that may be used to protect tables that contain confidential data. These methods have different impacts on the data utility of the published table. There may be disagreement among users over the utility of tabular data based on the different uses for the data. Will survey respondents report differently if they know that the agency will apply data swapping or add noise to the original values? How careful will respondents be to report accurately if they know the agency will be applying noise to their data? The level of statistical knowledge among data users and survey respondents varies and may be another factor that affects perceptions of applying alternative disclosure limitation methods.

More research is needed to study the perceptions of both data users and survey respondents concerning these data confidentiality issues. Accurately measuring the perceptions of EIA's user community and its survey respondents are important factors to consider in evaluating whether to apply a specific disclosure limitation method for protecting confidential data.

REFERENCES

- 1) Mayer, T.S. (2000). Privacy and confidentiality related research pertaining to the Census Bureau An annotated bibliography. Unpublished paper, Census Bureau. <http://www.census.gov/srd/papers/pdf/rsm2002-01.pdf>
- 2) Greenia, N., Jensen, J.B. & Lane, J. (2001). "Business Perceptions of Confidentiality," Chapter 16, p. 395, Confidentiality, Disclosure and Data Access, North-Holland (2001).
- 3) Statistical Working Paper No. 22 (Revised 2005). Federal Committee on Statistical Methodology. <http://www.fcsfm.gov/reports/#fcsfm>