

The Application of Protective Noise to Economic Programs

Richard A. Moore Jr., Paul B. Massell, and Jeremy M. Funk¹

Richard A. Moore, Jr.
Mathematical Statistician
Company Statistics Division

Census Advisory Committee of Professional Associations Meetings
October 18-19, 2007

Abstract

This paper discusses the application of protective noise for disclosure avoidance for several economic programs. Noise methodologies introduce small changes to the underlying microdata, providing protection to individual respondents without having to suppress a substantial number of cell values in published tables. Other techniques such as non-standard rounding and suppression of cells with small numbers of observations are used to provide added protection. The paper discusses (1) background on past and proposed disclosure techniques, (2) policy, privacy, and confidentiality requirements, (3) modifications for implementation, and (4) results of testing the procedure.

Questions:

- 1 How will noise affect data suppliers? In particular, does the ability to identify certain entities and to derive inexact values for them (see example on Page 4) alarm the data suppliers?
- 2 Is the typical data user satisfied with more data, even though the data are inexact?
- 3 For protection of the data, the Census Bureau's Disclosure Review Board does not allow us to provide users with the magnitude of the distortion introduced by noise. Is it acceptable to the Committee to provide data with slight distortions without providing information on how much the underlying microdata were distorted?
- 4 Trends may be affected by the application of noise. Which is preferable to the data user - a trend with an artificial spike or dip mostly attributed to the noise, or a time series with gaps caused by cell suppressions?

¹ Richard A. Moore, Jr. is a member of Company Statistics Division. Paul B. Massell and Jeremy M. Funk are members of the Statistical Research Division.

INTRODUCTION

The U. S. Census Bureau collects, tabulates, and releases economic data by various geographic, industry, and size class combinations. The final disseminated product is effectively a multi-dimensional matrix table with aggregate totals for each dimension. In accordance with U.S. Code, Title 13, Section 9, no data are published that would disclose the operations of an individual business.

For all economic programs, cell suppression has been used as the preferred disclosure risk avoidance technique. With cell suppression, the value of a cell is suppressed if that cell value could be used to obtain a close estimate for the value of one of the contributors of that cell. Cells suppressed in this fashion are known as primary suppressions. Additional cell values are suppressed so that the value of any primary suppression cannot be derived from the values of any of the remaining cells. These additional suppressions are known as complementary suppressions.

For many years cell suppression produced results with no disclosure breaches. For large multi-dimensional data products, however, this method can lead to a large number of suppressed cell values. To increase the utility of its data products, the Census Bureau has transferred one major program, the 2005 Nonemployer Statistics, from cell suppression to a multiplicative noise technique. It is also planning to transfer the Commodity Flow Survey, the Survey of Business Owners, and the Economic Census for Island Areas for the 2007 Economic Census cycle. Each program will use a slight variation of the EZS Noise method, which was developed at the Census Bureau in the 1990's by Evans, Zayatz, and Slanta (Journal of Official Statistics, Dec. 1998). The EZS method multiplies microdata values by factors slightly different than 1 (for example, 22 percent positive noise, multiplies each item by a factor of 1.22 (1.00+ 0.22), while 22 percent negative noise multiplies it by 0.78 (1.00 - 0.22). (22 percent is a hypothetical value for noise. For policy reasons, it does not necessarily reflect the true magnitude of the value of the noise which is used in these programs.) Most cell values compiled from aggregated “noisy” data differ by at most a few percentage points from the original estimates, yet the difference is usually sufficient to protect any individual respondent.

In publishing estimates compiled from noise-distorted values, the number of cell suppressions are greatly diminished. Table 1 in Attachment A shows a table from the 2004 Nonemployer Statistics program produced using the cell suppression technique. Most cells are suppressed in this table. Table 2 shows the analogous table for 2005 using noise instead of cell suppression. No cells are suppressed. Table 2 contains a complete set of inexact data. To warn the user, a one-sentence disclaimer in the header paragraph that reads, “...Values provided by each firm are slightly modified to protect the respondent's confidentiality. ...”

SATISFYING THE POLICY, PRIVACY, AND CONFIDENTIALITY REQUIREMENTS

Construction of the Noise Interval and Distribution for Random Assignment. The EZS method requires that the distribution from which the noise factors are selected must be

symmetric about 1, which ensures that the expected value of the average of any published cell (after the application of noise) is not biased; i.e., the expected value of the noise-distorted total for the cell is equal to the true value of the cell. The Census Bureau's Disclosure Avoidance Research Staff constructed the intervals from which the noise factors are to be selected as $(1-M-L, 1-M]$ and $[1+M, 1+M+L)$, where M is the minimum amount of recommended noise and L is the length of each interval. The values of M and L are determined by the Census Bureau's Disclosure Avoidance Research Staff and the Disclosure Review Board, and cannot be released to the public. Note that no noise factor can be selected from the interval $(1-M, 1+M)$. This guarantees that a prescribed level of protection is afforded to any single observation. Noise factors must differ between reporting entities, so that a data supplier, who may be able to estimate the amount of noise that was applied to his data, cannot use that factor to derive good approximations of the true value of data for individual suppliers in other cells. This method of assigning noise values is referred to "random assignment."

Balanced Hybrid Method. With random assignment, there is still a possibility that some cells may have values substantially different than the true values. For example, suppose a cell contains 10 observations. With random assignment, 9 of the 10 may receive positive noise. Some subject matter analysts wanted to avoid this for some key tables. Hence the Disclosure Avoidance Research Staff developed a "hybrid balancing assignment" for noise. With this method, a basic building block cell (for example, state by 3-digit NAICS) is identified; this is the summary level for which it is most important for noise-distorted cell values to be close to the true values. The microdata is sorted by building block cell and by descending size of reported value within each cell. Effectively the largest respondents in each cell are then randomly assigned a noise factor from the respective side of the noise distribution. The net amount of distortion that is added to the cell from these respondents is then calculated based on the factors and reported values, and recorded as a running total noise summand. If this summand is positive, then the next respondent in the cell will receive a negative noise direction, and visa versa. The magnitudes of all factors are draws from the "random" noise distribution, but the directions (or side of the distribution) are controlled to ensure noise cancellation. Although cell values may be more accurate for non-primary suppressions in the basic building block, other tables where the cells are defined by different variables (e.g., sales or employment size classes) will not benefit from this balancing.

Cell Suppression. Because some observations receive positive noise factors while others receive negative ones, there is a natural counteraction between observations. For cells with a large number of observations, this generally leads to a small amount of distortion in the final aggregated cell total. The application of noise for cells with one or two observations may identify a company and enough information that a data user may be able to derive an approximate value for that company (even though the derived value may be distorted by noise). A decision was made to suppress all information for cells with less than 3 responding entities. No values in other cells will be suppressed, so that it still may be possible for a data user to derive the data for some cells (see example below). It is assumed that users would check the documentation and realize that the value was suppressed due to a large amount of noise. It is also assumed that the data supplier would not be upset because the true data values have been distorted.

Example:	<u>NAICS</u>	<u># of Firms</u>	<u>Sales</u> <u>(\$1,000)</u>
	001	100	1,000
	0011	99	990
	0012	D	D
	002	200	10,000
	0021	198	9,000
	0022	D	D

MODIFICATIONS FOR IMPLEMENTATION

Non-standard rounding. EZS Noise is designed to protect individual respondents by adjusting their response values by small percentages. Unfortunately, standard rounding procedures (i.e, rounding fractional values to the nearest integer; e.g. 2.44 rounds to 2; note that in many cases (such as sales in the above example) the 2.44 and the 2 are additionally rounded to the nearest thousand) often remove the effects of noise on small response values. Specifically for small integer values, a small percentage change can produce a nearby real number which rounds back to the original integer value. We have recognized that in certain situations this occurs often and can have a significant affect on the level of protection afforded by noise. To deal with this issue we incorporate rounding methods that work to preserve the noise protection. The methods we use involve rounding record-level data prior to tabulation as well as some that are applied to table-level cell values.

Different methods are applied to the various surveys due to differences in data structure, sensitivity levels, and programming implementation requirements, although all are similar and produce similar results. In all of the methods, extra steps are taken to make sure that the noise is not entirely rounded out of small values. This may be done at the microdata record level to all or some selection of small records, or can be performed at the table level if table additivity is not required. These additions to the noise procedure generally have little overall effect on the method from a data quality standpoint, since only the smallest values are affected.

RESULTS

The multiplicative noise technique (with the enhanced modifications) allows the Census Bureau to provide more data to its user, even though the data products provided are based on distorted microdata. Cells, which are not primary suppressions, should contain values with relatively small amounts of net distortion, while primary suppression cells contain values that generally contain higher amounts of distortion. For each program, the table below shows (1) the percentage of cells suppressed by each method, (2) the percentage of primary suppressions under the cell suppression method, and (3) the amount distortion for 75 percent of the cell values.

Research Results Using Noise

Program and Year Used to Evaluate Method	Percentage of All Cells With Published Values		Percentage of All Cells That Are Primary Suppressions	Distortion Range for the Least Distorted Cells (Lowest 75%)
	Cell Supp	Noise Method		
Nonemployer Statistics (2004)	38%	73%	33%	0 to 5%
Economic Census of Island Areas (2002)	67%	80%	24%	0 to 5%
Survey of Business Owners (2002)	67%	93%	7%	0 to 2%
Commodity Flow Survey (2002)	NA	NA	< 1%	0 to 2%

Values of cells that are primary suppressions require more distortion to protect than non-primary cells. About 33 percent of the of the cells in the 2004 Nonemployer Statistics and about 24 percent of those in the 2002 Economic Census of Island Areas are primary suppressions under the traditional cell suppression rule. The high percentage of primary suppressions raise the upper bound of the distortion range (shown in the last column in the table above) to 5 percent for these programs. Tables for programs such as the 2002 Survey of Business Owners and the 2002 Commodity Flow Survey contained a small percentage of primary suppressions. Consequently, 75 percent of cell values were distorted by 2 percent or less.

**Table 1. 2004 Nonemployer Statistics
Accommodations and Food Service Sector (NAICS 72)
Calvert County, MD
Using a Cell Suppression Technique**

Data based on the 2004 Nonemployer Statistics. Includes only firms subject to federal income tax. [Introductory text](#) includes scope and methodology, non-sampling error, and confidentiality protection.

NAICS	Description	Firms	Receipts (\$1,000)
72	Accommodations and Food Service	44	1,339
721	Accommodations	D	D
7211	Traveler Accommodations	D	D
7213	Rooming And Boarding Houses	D	D
722	Food Services and Drinking Places	D	D
7221	Full Service Restaurants	D	D
7222	Limited Service Eating Places	D	D
7223	Special Food Services	28	292

Note: Table 1 is based on the 2004 data, while Table 2 is based on 2005 data. This accounts for the difference in firm counts and receipts throughout the table (e.g., at the 2-digit sector level).

**Table 2. 2005 Nonemployer Statistics
Accommodations and Food Service Sector (NAICS 72)
Calvert County, MD
Using a Multiplicative Noise Technique**

Nonemployer Statistics originate from tax return information of the Internal Revenue Service. The 2004 and 2005 Nonemployer totals may be low due to late tax reporting in hurricane impacted counties/regions. The data are subject to nonsampling error such as errors of self-classification by industry on tax forms, as well as errors of response, nonreporting and coverage. Values provided by each firm are slightly modified to protect the respondent's confidentiality. For further information about methodology and data limitations, see [Survey Methodology](#).

NAICS	Description	Firms	Receipts (\$1,000)
72	Accommodations and Food Service	56	2,523
721	Accommodations	12	263
7211	Traveler Accommodations	9	254
7213	Rooming And Boarding Houses	3	9
722	Food Services and Drinking Places	44	2,260
7221	Full Service Restaurants	3	1,233
7222	Limited Service Eating Places	7	433
7223	Special Food Services	34	594