Journal of Official Statistics, Vol 14, No. 4, 1998, pp. 537–551

Using Noise for Disclosure Limitation of Establishment Tabular Data

Timothy Evans, Laura Zayatz, and John Slanta[1]

We propose a new disclosure limitation method for establishment magnitude tabular data in which noise is added to the underlying microdata prior to tabulation. The proposed method has several advantages compared to the standard method of cell suppression: it enables some information to be provided within more cells of the table, it eliminates the need to coordinate cell suppression patterns between tables, and it is a much less complicated and time-consuming procedure than cell suppression. In this article we outline the proposed procedure for adding noise to the underlying establishment microdata, discuss the advantages and disadvantages of adding noise as compared to cell suppression, and describe the results of using noise with data from one survey.

Key words: Disclosure limitation; noise; confidentiality; cell suppression; magnitude data.

1. Introduction

The responding unit in many economic surveys and censuses conducted by statistical agencies is the establishment. Individual establishments' responses are weighted (where appropriate) and estimates of quantities of interest such as value of shipments are generally produced by categorical variables like Standard Industrial Classification (SIC) code and geography. The categorical variables define a table (for example the rows might be SIC codes and the columns might be geographic areas). Then the ''quantity of interest'' is aggregated over all units of analysis in each cell. Such tables are called tables of magnitude data. Given the geographic information and other characteristics on which tables are based, in conjunction with common knowledge and publicly available sources, it is generally a reasonable assumption that the set of establishments contributing to a cell in such a table is well-known to data users.

Many statistical agencies collect information promising that all responses will be held confidential (Duncan, Jabine, and de Wolf 1993). Those same agencies attempt to release as much statistically valid and useful data as possible without violating the confidentiality pledge. Techniques used to protect data confidentiality are called ''disclosure limitation'' procedures (see Federal Committee on Statistical Methodology 1994 for a review of disclosure limitation methodologies for all types of data and for an annotated bibliography of the literature on disclosure limitation). The disclosure limitation procedures for magnitude data are designed to prevent data users from being able to recover any respondent's reported values using values appearing in the published tables. A statistical agency

[1] Bureau of the Census, Washington DC, U.S.A.

must ensure that a cell value does not closely approximate data for any one respondent in the cell and, moreover, that one respondent or a coalition of respondents cannot subtract their contribution(s) from the cell value to achieve a ''close'' estimate of the contribution of another respondent (Cox and Zayatz, 1993; Dalenius, 1982).

2. Cell Suppression

The current widely accepted disclosure limitation technique used for establishment magnitude tabular data is cell suppression. Cox (1992) and Sande (1984) discuss cell suppression methodologies. Cells that pose a disclosure risk are typically identified using one of two rules – the *n-k* rule or the *p%* rule (see Federal Committee on Statistical Methodology 1994 for a detailed explanation of these rules). All cells that fail the disclosure rule are called sensitive cells. In the context of cell suppression, these cells are also often called primary suppressions.

Cell suppression limits disclosure by removing from publication (suppressing) all sensitive cells plus sufficiently many additional cells, called complementary suppressions, to ensure that the values of the primary suppressions cannot be narrowly estimated through manipulation of additive relationships between cell values and totals (Cox and Zayatz 1993). When a cell is suppressed, its total value is removed from the cell and replaced with some type of symbol indicating that the value is withheld to prevent disclosure.

While the concepts behind determining whether a particular cell is a disclosure risk are relatively simple, the process of choosing complementary suppressions to protect these sensitive cells is quite complicated. The methodology by which complementary suppressions are chosen, as well as the accompanying computer software, is very difficult to understand for anyone without a background in linear programming. Because of the structure of the computer programs, often the process must be performed separately for each data product. Among other things, this means that agency staff must keep track, from one data product to the next, of which cells have previously been suppressed (and hence must be suppressed and protected in all subsequent data products) and which cells have previously been published (and hence cannot be used as complementary suppressions).

Coordinating suppression patterns among tables becomes impracticable in the presence of the multiple requests for special tabulations which frequently follow standard publications. To truly prevent any disclosures, an agency must keep track of *all* special tabulations requested by *all* data users, identifying not only those cells that were suppressed in any of the tabulations but also any unsuppressed cells that could be used in conjunction with unsuppressed cells from another tabulation to recover the value of a suppressed cell. Thus the agency must keep an ongoing record of all interrelationships between all cells across all publication tables and special tabulations, a programming nightmare. Many agencies simply do not have the resources to do this.

Another major drawback of cell suppression is that it suppresses much information that is not at risk of disclosure. Any cell that is used as a complementary suppression but that is not itself a primary suppression represents information that could have been published if there were some other way of protecting the sensitive cells. Particularly at fine levels of detail, including most special tabulations, the need for complementary suppressions often results in tables full of suppressed cells.

3. Introduction of Noise to Microdata Prior to Tabulation

3.1. General description

We propose an alternative method of protecting individual respondents. The method involves adding noise to their data. This procedure should not be confused with noise procedures aimed at protecting and releasing public use establishment microdata (McGuckin and Nguyen 1990). This procedure is for establishment magnitude tabular data. We propose perturbing each responding establishment's data by a small amount, say 10% (the per cent to remain confidential within the statistical agency). Then if a cell contains only one establishment, or if a single establishment dominates the cell, the value in the cell will not be a close approximation to the dominant establishment's value because that value has had noise added to it (in this case, it has been changed by about 10%). By adding noise, we avoid disclosing the dominant establishment's true value.

To each establishment in our sample or census we assign a multiplier, or noise factor. Then all establishments have their values multiplied by their corresponding noise factors before the data are tabulated. Note that because the same multiplier is used with an establishment wherever that establishment is tabulated, values are consistent from one table to another. That is, if the same cell appears on more than one table, it has the same value on all tables.

Note that we add noise to each establishment prior to any tabulations. This is *not* the same as attempting to add noise on a cell-by-cell basis. We rely on a random assignment of the multipliers to control the effects of the noise on different types of cells. The noise should have its greatest effect on sensitive cells, while the effect of the noise on cells that would not be primary suppressions should be minimal. Thus we aim to protect individual establishments without compromising the quality of our non-sensitive estimates.

3.2. The multipliers

For purposes of illustration, let us assume we want to introduce roughly 10% noise into each establishment's values. The actual percentage used by a statistical agency would be confidential. To perturb an establishment's data by about 10%, we multiply its data by a number that is close to either 1.1 or 0.9. We could use any of several types of distributions from which to choose our multipliers, and the distributions would remain confidential within the agency. For our example, to perturb an establishment's data in a positive direction, we could choose the multipliers from a normal distribution with mean 1.1 and a small variance, perhaps .05 or .01. In any case, we want to use a distribution centered at or near 1.1 and with small variance. If we want to ensure that all multipliers perturbing data in a positive direction are at least 1.1, guaranteeing at least a 10% change in value to single-establishment cells, we can simply truncate our distribution at 1.1 and discard the portion below 1.1.

Whatever distribution we decide to use for generating multipliers near 1.1, it is of paramount importance that we use the same shape distribution, or rather its ''mirror image,'' to generate multipliers near 0.9. In other words, if we consider the two distributions together, the overall distribution of the multipliers should be symmetric about 1. The reason for this condition is discussed in Section 3.3.

Under current practices, the unit of analysis for disclosure limitation is the *company*. That is, we seek to protect respondent data at the company level as well as for individual establishments within the company. Because company-level values must be protected, all noise for a single company should either inflate or deflate that company's true values. In other words, all establishments from the same company should be perturbed in the same direction and hence have approximately the same multiplier. This way, if all of the establishments contributing to a cell belonged to the same company, the resulting cell estimate would be perturbed by, for our example, about 10%. Otherwise, the cell estimate could be very close to the company's true value if the noise in the positively-perturbed establishments (multipliers > 1) and the noise in the negatively-perturbed ones (multipliers < 1) happened to roughly cancel each other out. Thus by perturbing all of a company's establishments in the same direction, we ensure that company-level data is protected.

3.3. *Assignment of multipliers and its effect on estimates*

We want to assign the multipliers in such a way that we minimize the effect of the noise on those cells that are not at risk of disclosure. In particular, cell values at higher levels of aggregation are not generally sensitive, and we would like these values to contain as little noise as possible. In this section, we will concern ourselves with data from a census. The next section (3.4) extends the methodology to survey data.

We begin by randomly assigning each responding company a *direction* of perturbation. Using our example with 10% as our base for perturbation, this is equivalent to determining if all establishments in that company will have multipliers close to 1.1 or close to 0.9. We then randomly assign a multiplier to each establishment 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 potential assignments is as follows:

Example 1:

Company	*Establishment*	*Direction*	*Multiplier*
Company A		1.1	
	Establishment A1		1.12
	Establishment A2		1.09
	Establishment A3		1.10
	Establishment A4		1.11
Company B		0.9	
	Establishment B1		0.89
	Establishment B2		0.93
Company C		1.1	
	Establishment C1		1.08

Intuitively, the expected value of the amount of noise present in any cell value is zero, due to the symmetry of the distribution of the multipliers and the random assignment of direction of perturbation and multipliers within the companies. The probability that a company's establishments 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 establishment is 0, and the amount of noise in any cell value is simply the sum of the noise in its component establishments. Let Y be the noise-free cell value and Y_N be the noise-added cell value. Thus for establishment i in the cell j, we have:

$$Y = \sum_{i \in j} value_i$$

and

$$Y_N = \sum_{i \in j} (multiplier_i \times value_i)$$

Let

$$e = Y - Y_N$$

Given that

$$E(multiplier_i) = 1$$

we have

$$\begin{aligned} E(Y_N) &= E\left(\sum_{i \in j} multiplier_i \times value_i\right) \\ &= \sum_{i \in j} E(multiplier_i \times value_i) \\ &= \sum_{i \in j} (value_i \times E(multiplier_i)) \\ &= \sum_{i \in j} (value_i \times 1) \\ &= \sum_{i \in j} value_i \\ &= Y \end{aligned}$$

and hence

$$E(e) = 0$$

Thus the noise procedure does not introduce any consistent bias into the cell values.

For non-sensitive cells, values are not altered a great deal, as we will see in Section 4. For these cells, the establishments that are perturbed in the positive direction and those that are perturbed in the negative direction will generally balance each other out. In contrast, a cell that is dominated by a single contributor will most likely contain a large amount of noise. If the largest contributor is very large compared to all others in the cell, it is much less likely that positively-perturbed establishments and negatively-perturbed establishments will cancel each other out when determining the amount of noise present in the cell value. Looked at another way, the more dominant the largest contributor, the more the amount of noise present in the cell value will resemble the amount of noise present in the largest contributor (about 10%). Thus the cells that are at greatest risk of disclosure in general receive the most noise, and the noise in the cell total prevents users from being able to recover an individual respondent's true value from the published value. This is illustrated in the following examples:

Example 2:

Sensitive cell

	True establishment values	*Noise-added establishment values*
	10,000	10,000 × 1.11 = 11,100
	300	300 × 0.89 = 267
	200	200 × 1.12 = 224
Cell total	10,500	11,591

Note that in this example of a sensitive cell, the cell value is changed by (11,591–10,500)* 100/10,500 = 10.39%.

Example 3:

Non-sensitive cell

	True establishment values	*Noise-added establishment values*
	10,000	10,000 × 1.11 = 11,100
	8,000	8,000 × 0.89 = 7,120
	5,000	5,000 × 1.12 = 5,600
Cell total	23,000	23,820

In this example of a non-sensitive cell, the cell value is changed by (23,820–23,000) * 100/23,000 = 3.57%.

3.4. Adding noise to census data versus sample survey data

When generating tables from census data using the noise approach for disclosure limitation, a responding establishment's contribution to a cell simply becomes

establishment value × multiplier.

The examples in Section 3.3 show how noise would be applied and cells would be tabulated for census data.

In sample surveys, each respondent's data is generally weighted inversely proportional to the establishment's probability of being included in the sample. For establishments with large weights, the weight itself offers some protection against disclosing the respondent's actual reported values (Willenborg and de Waal 1998). For sample data, to reflect the protection already provided by the sample weight, noise is applied as follows:
For each establishment in a cell, the establishment's contribution to the cell becomes

establishment value × [multiplier + (weight − 1)]

These noise-added establishment contributions are then added to obtain total cell value. Note that noise is added only to one multiple of each establishment's value, and the remaining (weight − 1) multiples, which conceptually represent the contributions of other unsampled establishments, have no noise added. Also notice that as the weight approaches 1, i.e., as the establishment comes closer and closer to representing only itself, the formula degenerates to the census case, as the following example illustrates.

Example 4:

True establishment contributions		*Noise-added establishment contributions*	*Per cent*
Value × *Weight*		*Value* × [*Multiplier* + (*Weight* − 1)]	*Difference*
	10,000 × 1 = 10,000	10,000 × [0.89 + (1 − 1)] = 8,900	11.00
	800 × 5 = 4,000	800 × [1.12 + (5 − 1)] = 4,096	2.40
	500 × 7 = 3,500	500 × [0.91 + (7 − 1)] = 3,455	1.29
Cell total	17,500	16,451	5.99

This procedure has the effect of changing contributions for certainty or near-certainty establishments (those having weights close to or equal to 1) by a large amount while changing contributions for establishments with large weights by a small amount. This is desirable because we are more concerned with the disclosure risk of certainty establishments, whose values are not protected by their weights.

Let $\hat{Y}$ be the noise-free cell estimate and $\hat{Y}_N$ be the noise-added cell estimate. Thus for establishment i in cell j, we have:

$$\hat{Y} = \sum_{i \in j} value_i \times weight_i$$

and

$$\hat{Y}_N = \sum_{i \in j} [(multiplier_i + weight_i - 1) \times value_i]$$

Let

$$e = \hat{Y} - \hat{Y}_N$$

Given that

$$E(multiplier_i) = 1$$

we have

$$E(\hat{Y}_N) = E\left(\sum_{i \in j} (multiplier_i + weight_i - 1)\ value_i\ t_i \right)$$

where $t_i = 1$ *if* i^{th} *sampling unit is selected*
$= 0$ *otherwise*

$$\begin{aligned} E(\hat{Y}_N) &= \sum_{t=0} E\left(\sum_{i\in j}(\mathit{multiplier}_i + \mathit{weight}_i - 1)\ \mathit{value}_i\ t_i \mid t_i = t_i\right)\ P(t_i = t_i) \\ &= E(\sum_{i\in j}(\mathit{multiplier}_i + \mathit{weight}_i - 1)\ \mathit{value}_i)\ \pi_i \\ &= \sum_{i\in j}(\mathit{value}_i\ \pi_i\ E(\mathit{multiplier}_i) + (\mathit{weight}_i - 1)\ \mathit{value}_i\ \pi_i) \\ &= Y \end{aligned}$$

Since
$E(\hat{Y}) = Y$

we have

$E(\hat{Y}_N) = E(\hat{Y})$
and therefore
$E(e) = 0$

Thus again the noise procedure does not introduce any consistent bias into the cell values. Note also that we have the following properties:

$$COV(\hat{Y}, e) = 0$$

and

$$\begin{aligned} \sigma^2(\hat{Y}_N) &= \sigma^2(\hat{Y} + e) \\ &= \sigma^2(\hat{Y}) + \sigma^2(e) + 2\ COV(\hat{Y}, e) \\ &= \sigma^2(\hat{Y}) + \sigma^2(e) + 0 \\ &= \sigma^2(\hat{Y}) + E(e^2) - (E(e))^2 \\ &= \sigma^2(\hat{Y}) + E(e^2) \end{aligned}$$

Thus the introduction of noise causes the variance of an estimate to increase by an amount equal to the square of the noise term. An unbiased estimator of the variance is:

$$\hat{\sigma}^2(\hat{Y}_N) = \hat{\sigma}^2(\hat{Y}) + e^2 = \hat{\sigma}^2(\hat{Y}) + (\hat{Y}_N - \hat{Y})^2$$

3.5. *Flagging cells with a large amount of noise*

The percentage of noise in a cell is defined as the per cent by which the noise-added value for the cell differs from the true noise-free value. Thus we have to calculate both the noise-added and noise-free values for each cell in order to quantify the amount of noise each cell contains. All resulting table cells containing a large percentage of noise, say a 7% change in value or more, (the chosen percentage would be confidential), are flagged so users would know that the values may not be useful. This set of cells will encompass most sensitive cells, as well as a few non-sensitive cells that received a lot of noise simply through randomness. The description of the flag explains how and why noise was added and lets users know that disclosure limitation has been performed. Users (we hope) are thus discouraged from using the inaccurate, noise-added totals in sensitive cells to try to recover a value for any individual contributor.

We also use the same flag on any cells that were identified as sensitive (i.e., failed the disclosure rule) before noise was added but that, because of randomness of multipliers, did not exceed our noise threshold (7% in our example). In this case, users at least *think* the cell contains a lot of noise and hesitate to treat the cell value as reliable. We expect relatively few cells of this type.

Cells exceeding the noise threshold, as well as sensitive cells not sufficiently protected by the noise, will contain a flag but no published value. The value of the cell may still be derivable, but the fact that the value did not actually appear in the cell would draw attention to the fact that we did not consider the estimate reliable. This is similar to how we treat cells having high coefficients of variation (CVs) in survey publications. By not publishing actual values, we may also lessen the *appearance* of disclosure for single-establishment cells and for sensitive cells that did not receive much noise.

4. Results with Actual Survey Data

To get an idea of how well the noise technique actually works in practice, we tested it with data from the U.S. Bureau of the Census' Research and Development (R and D) Survey, a survey of companies' research and development expenses. In this survey, estimates of R and D expenses are computed for 26 SICs or SIC groupings, and within each SIC the expenses are separated into corporate-sponsored R and D and federally-sponsored R and D.

We randomly assigned responding companies a direction of perturbation. To then generate the establishment multipliers, we experimented with several distributions. We tried the following options:

1) normal distributions, centered at 0.9 and 1.1, respectively, and with small standard deviation $\sigma = .02$ —$X \sim$ N(0.9, 0.02) and $X \sim$ N(1.1, 0.02)
2) truncated normal, using the same distributions as in (1) but discarding any number between 0.9 and 1.1
3) ''ramp'' distributions — modes at 1.1 and 0.9, $f(x) = 0$ between 0.9 and 1.1, $f(x) = 0$ at 1.2 and 0.8, and $f(x)$ inversely proportional to $(x - 1.1)$ between 1.1 and 1.2 and inversely proportional to $(0.9 - x)$ between 0.8 and 0.9
4) scaled Beta distributions — X ~ .1 B(6,2) + 0.8 and X ~ .1 B(2,6) + 1.1

After generating a multiplier for an establishment, we applied the multiplier to the establishment's data items using the formula in Section 3.4 for sample data.

We reproduced a table from the R and D publication, which shows R and D expenses broken out by federally-sponsored vs corporate-sponsored, for the 26 SIC groupings. (Table 4.1 below shows the structure of the R and D table.) We ran 100 simulations of the R and D table for each of the four options described above for generating multipliers, and we compared each option to the original noise-free table, looking at the per cents by which the cell values changed as a result of the noise. We could find no consistent differences among the four options, except perhaps that the ''ramp'' distribution produced larger variability in the amount of noise present in a cell. Since all four options seemed to perform satisfactorily, we chose to use the Beta distribution. The ability to control the location and relative height of the mode of the distribution and the fact that the distribution can be scaled to fit into any finite interval seemed desirable qualities.

Next we ran 1,000 replications of the R and D table, using the Beta distribution to generate multipliers, and computed summary statistics to describe the behavior of the cells over all replications. Below is a copy of the table showing, for each cell, the ratio of a) the average of the 1,000 noise-added values of the cell to b) the true noise-free value. Thus if there is no tendency for the noise to change the value of a cell in any particular direction, the values in the table should be close to 1, i.e., the average noise-added value for any cell should be close to the true cell value. A dash (-) indicates that the cell has a value of zero. The SIC groupings are shown as stub numbers 1 through 26 and do not appear in the same order as in the R and D publication. Sensitive cells are underlined.

Note that the values are indeed close to 1, for both sensitive and non-sensitive cells. The largest and smallest values are, respectively, 1.00326 and 0.99692. It is clear that the symmetry of the distribution of the multipliers and the randomness of the direction

Table 4.1. Ratio of noise-added value to noise-free value.
Average noise-added value over 1,000 simulations of R and D table, divided by true unperturbed value

stub #	total R and D	federal	company
1	0.99914	0.99874	0.99915
2	0.99932	0.99761	0.99932
3	0.99955	0.99725	0.99966
4	1.00128	1.00152	1.00127
5	1.00153	(-)	1.00153
6	0.99895	1.00300	0.99889
7	1.00091	1.00294	1.00084
8	0.99955	0.99837	0.99970
9	0.99996	1.00212	0.99978
10	1.00017	(-)	1.00017
11	0.99950	1.00326	0.99950
12	1.00082	0.99711	1.00082
13	0.99945	0.99738	0.99989
14	1.00049	1.00047	1.00050
15	0.99900	0.99757	1.00013
16	1.00028	0.99983	1.00029
17	0.99882	0.99737	0.99960
18	0.99900	0.99961	0.99863
19	0.99956	0.99968	0.99952
20	1.00069	(-)	1.00069
21	0.99773	0.99762	0.99776
22	0.99946	0.99692	0.99987
23	0.99993	0.99892	1.00024
24	0.99984	(-)	0.99984
25	1.00100	1.00234	1.00033
26	0.99925	0.99832	0.99936
Total	0.99949	0.99944	0.99950

of perturbation ensure that the expected value of the noise present in any estimate is zero (i.e., the expected value of the ratio of the noise-added value to the noise-free value is 1). Hence the noise does not introduce any bias into the estimates, as was shown in Section 3.3.

Note that while the expected value of the amount of noise in any one *establishment* is also zero (since the symmetry of the distribution of multipliers implies that the expected value of any particular multiplier is 1), in practice this will not happen because of the bimodality of the distribution; a multiplier can never actually equal 1. In fact, in the degenerate case where an estimate is composed of only one establishment, the estimate is guaranteed to contain at least 10% noise.

To get an idea of how much noise would typically be present in a cell after a *single* application of the noise, we looked at the standard deviation of the 1,000 noise-added observations in each cell. We standardized these by dividing by the true cell value. If we consider the true value of the cell estimate $\hat{Y}$ to be ''fixed'' for purposes of adding noise, then the standard deviation of the noise-added values $\hat{Y}_{\mathrm{N}}$ is simply the standard deviation of the noise itself: writing $\hat{Y}_{\mathrm{N}} = \hat{Y} + e$ and taking $\hat{\sigma}(\hat{Y})$ and $\mathrm{Cov}(\hat{Y},\mathrm{e})$ to be zero, we have $\hat{\sigma}(\hat{Y}_N|\hat{Y}) = \hat{\sigma}(e)$. The value in the table, $\hat{\sigma}(\hat{Y}_N|\hat{Y})/\hat{Y}$, can be thought of as the coefficient of variation (CV) of the noise-added estimate, given the noise-free estimate, i.e., $\mathrm{CV}(\hat{Y}_{\mathrm{N}}|\hat{Y})$. Table 4.2 below shows these conditional ''CVs''over the 1,000 replications. Again, sensitive cells are underlined.

Note that the conditional CVs are generally much higher in the sensitive cells than in the non-sensitive ones. The variability of the amount of noise present in sensitive cells is much larger, so a sensitive cell should be much more likely than a non-sensitive cell to contain a large amount of noise after a single application of the noise procedure. This is exactly what we want, since it is the sensitive cells whose values need to be protected.

To confirm this idea, we looked at the amount of noise that was typically present in different types of cell. For each non-zero cell, we computed the *absolute* per cent noise present in the cell for each replication. We then computed an overall ''per cent noise'' by averaging these absolute percentages over all 1,000 replications. (Note that if we did not use the absolute value of the percentage, the average over all replications would be close to zero and would tell us nothing about the typical behavior of the cell.) Then we looked at the distribution of this ''per cent noise'' variable among cells of various types. Table 4.3 below gives the results.

The distinction between marginal cells and interior cells shows that interior cells on average received more noise. This is a desirable result, since interior cells are composed of fewer establishments and are more likely to be sensitive. The noise technique appears to leave marginal estimates with relatively little noise, roughly between 2 and 3 per cent.

Cells that would have been primary suppressions receive noticeably more noise than non-sensitive cells. Again, this is what we want, because these are the cells whose values need to be protected. Complementary suppressions are shown separately to illustrate the fact that the noise technique would allow these cells to be published with relatively little noise, thus providing data users with more information than would have been the case with cell suppression.

Because of the element of randomness in assigning multipliers, we do not expect *all* sensitive cells to receive a lot of noise (see Section 3.7), nor do we expect that none of

Table 4.2. Conditional ''CVs'' of noise-added values.
Standard deviation of the 1,000 simulations, divided by the true noise-free estimte

stub #	total R and D	federal	company
1	0.03435	0.04836	0.03426
2	0.03344	0.12550	0.03335
3	0.01512	0.12511	0.01001
4	0.04448	0.05300	0.04414
5	0.06648	(-)	0.06648
6	0.03719	0.12121	0.03959
7	0.03875	0.12505	0.03586
8	0.02008	0.07398	0.01349
9	0.00294	0.10999	0.00747
10	0.01265	(-)	0.01265
11	0.02395	0.12604	0.02399
12	0.03213	0.12507	0.03246
13	0.02200	0.11817	0.00470
14	0.01596	0.01794	0.01589
15	0.04755	0.10916	0.00259
16	0.00937	0.01394	0.00961
17	0.04861	0.10957	0.01592
18	0.03150	0.00757	0.04686
19	0.01954	0.02110	0.01922
20	0.03700	(-)	0.03700
21	0.08972	0.09229	0.08896
22	0.01880	0.11383	0.00473
23	0.00324	0.04377	0.00979
24	0.00367	(-)	0.00367
25	0.04369	0.10209	0.01500
26	0.03716	0.07130	0.03320
Total	0.01912	0.01843	0.01931

the non-sensitive cells will receive a lot of noise. Table 4.4 below gives the breakdown, by type of cell, of which of the 77 nonzero cells in our test table received a lot of noise (where we define ''a lot'' as at least 7%) and which did not receive much.

This table further illustrates that the noise technique generally leaves non-sensitive cells (including marginal totals) relatively noise-free, while most sensitive cells receive a lot of noise. The few sensitive cells that do not exceed the noise threshold would be flagged as described in Section 3.5, along with both sensitive and non-sensitive cells that do exceed it.

5. Conclusions

Adding noise to establishment-level data before producing tables has several advantages over the traditional cell suppression techniques. First, it is a far simpler and less time-consuming procedure than cell suppression. Computer programs for adding noise are much easier to write, modify, run, and understand than the programs that currently exist for choosing cell suppression patterns.

Another important advantage of adding noise is that it eliminates the need to coordinate

Table 4.3. Amount of noise in non-zero cells, by type of cell

% noise in:	amount of noise			
	avg	median	max	min
marginal cells (29)	2.88	2.36	8.89	0.24
interior cells (48)	5.18	3.60	12.52	0.21
cells that would have been primary suppressions (11)	11.11	12.08	12.52	5.19
cells that would have been complementary suppressions (12)	2.77	3.24	4.73	0.24
unsuppressed cells (54)	3.27	2.00	11.32	0.21
all (nonempty) cells (77)	4.32	3.31	12.52	0.21

cell suppressions between tables. Under the current cell suppression practices, disclosure analysis involves keeping track, from one data product to another, of all cells that have previously been published and all cells that have previously been suppressed. Keeping track of suppressions is difficult to orchestrate and difficult to understand. However, using noise to protect estimates would make this unnecessary.

Also, with cell suppression, users lose information both for cells which are primary suppressions and for those that are complementary suppressions. With the noise technique, sensitive cells (those that would normally be primary suppressions) would in general contain a lot of noise and be flagged as such. In contrast, non-sensitive cells would end up with little noise, including most of the non-sensitive cells that would have been used as complementary suppressions. Thus for publications which normally contain many complementary suppressions, the noise technique should provide data users with more valuable information.

But what about protection? With cell suppression, although actual values are suppressed, data users can use linear programming techniques to calculate a possible range for each suppressed value. Statistical agencies assign primary and complementary suppressions to ensure that a respondent's value cannot be closely estimated (ranges must meet size requirements). With noise, data users may be able to obtain a point estimate that can be associated with a given respondent, but this estimate would contain ''a lot'' of noise (statistical agencies would determine the amount they feel comfortable with just as they determine the range size requirements). Some may argue that the added noise does not provide enough protection to values in single-establishment cells. Under the cell

Table 4.4. Counts of cells having large vs small amounts of noise

type of cell	$\lvert$ noise $\rvert \geq 7\%$	$\lvert$ noise $\rvert < 7\%$
sensitive (11)	10	1
non-sensitive (66):	7	59
complementaries (12)	0	12
unsuppressed (54)	7	47
marginal (29)	1	28
interior (37)	6	31
total (77)	17	60

suppression approach, if a cell has only a single establishment contributing to it, the cell's value would be suppressed and the cell would simply contain a ''D.'' Using the noise technique, the cell would contain a flag noting that the value in the cell had been severely altered, but the actual value may still be derivable using other cells in the same row or column. The flag may lessen the *appearance* of disclosure, since no value would appear in the cell. However, the respondent may still feel uneasy about the derived number seeming to be an estimate of his actual value, even if the estimate contains a lot of noise and is flagged as being unreliable. The suppression approach may give the appearance of offering more protection.

It is possible that some respondents may resent putting time into preparing good responses if they know the statistical agency is going to add noise to them. We need to emphasize that noise would be added in an unbiased way so as to preserve the statistical properties of the data while having a negligible effect on non-sensitive estimates.

Also there may be concern on the part of some data users as to the quality of the data after noise has been introduced. The users' desire for multiple special tabulations and their desire to see more published cells (at the expense of noise) should be weighted against their desire for true values (at the expense of suppressions).

The results of our test with the R and D Survey indicate that the idea of adding noise as a disclosure limitation strategy warrants further consideration. We have thus far been concerned with the effect of noise on the behavior of the level estimates in our published tables, and in this regard it performs well. Under our scheme for assigning multipliers, the noise does not appear to introduce any bias into the estimates. We have also shown that, in general, sensitive cells end up containing larger amounts of noise than non-sensitive cells; thus noise provides protection where it is most needed. See Evans, Zayatz, and Slanta (1996) for ideas on the use of sorting and raking to better preserve non-sensitive estimates. Looking beyond the behavior of level estimates, further research is required to investigate the effect of noise on trend estimates, longitudinal studies, inter-variable relationships, and other types of analysis that data users typically perform with the published estimates (Evans, Zayatz, and Slanta 1996; Evans 1997).

The noise technique is probably not suitable for all data products; some surveys publish data at such levels of aggregation that disclosure is not an issue. However, for surveys in which cell suppression currently creates problems, the prospects are encouraging. In light of our results and the flexibility and simplicity that the noise technique offers, the addition of noise to the underlying microdata could become a viable alternative to cell suppression for disclosure avoidance with establishment tabular data.

6. References

Cox, L.H. (1992). Solving Confidentiality Protection Problems in Tabulations Using Network Optimization: A Network Model for Cell Suppression in U.S. Economic Censuses. Proceedings of the International Seminar on Statistical Confidentiality, International Statistical Institute, Dublin, 229–245.

Cox, L.H. and Zayatz, L. (1993). Setting an Agenda for Research in the Federal Statistical System: Needs for Statistical Disclosure Limitation Procedures. Proceedings of the Section on Government Statistics, American Statistical Association, 121–126.

Dalenius, T. (1982). Disclosure Control of Magnitude Data. Statistisk tidskrift, No. 3, 173–175.

Duncan, G.T., Jabine, T.N., and de Wolf, V.A. (eds.) (1993). Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics. Committee on National Statistics and the Social Science Research Council, National Academy Press, Washington, DC.

Evans, B.T. (1997). Effects on Trend Statistics of the Use of Multiplicative Noise for Disclosure Limitation. Proceedings of the Section on Government Statistics, American Statistical Association, to appear.

Evans, B.T., Zayatz, L., and Slanta, J. (1996). Using Noise for Disclosure Limitation of Establishment Tabular Data. Proceedings of the Annual Research Conference, U.S. Bureau of the Census, Washington, DC 20233, 65–86.

Federal Committee on Statistical Methodology (1994). Report on Statistical Disclosure Limitation Methodology. Statistical Policy Working Paper 22, U.S. Office of Management and Budget, Washington, DC.

McGuckin, R.H. and Nguyen, S.V. (1990). Public Use Microdata: Disclosure and Usefulness. Journal of Economic and Social Measurement, 16, 19–39.

Sande, G. (1984). Automated Cell Suppression to Preserve Confidentiality of Business Statistics. Statistical Journal of the United Nations, ECE 2, 33–41.

Willenborg, L. and de Waal, T. (1998). Statistical Disclosure Control and Sampling Weights. Journal of Official Statistics, 13, 417–434.

Received August 1997
Revised August 1998