USING NOISE FOR DISCLOSURE LIMITATION OF ESTABLISHMENT TABULAR DATA

Timothy Evans, Laura Zayatz, and John Slanta
Bureau of the Census

ABSTRACT

The Bureau of the Census is looking into new methods of disclosure limitation for use with establishment tabular data. Currently we use a strategy that suppresses a cell in a table if the publication of that cell could potentially lead to the disclosure of an individual respondent's data. As an alternative to cell suppression that would allow us to publish more data and to fulfill more requests for special tabulations, we are experimenting with adding noise to our underlying microdata. By perturbing each respondent's data, we can provide protection to individual respondents without having to suppress cell totals.

While adding noise is a much less complicated and time-consuming procedure than cell suppression, the question remains as to the utility of the data after noise is added. To preserve the quality of aggregate estimates that would not normally be at risk for disclosure, we tested the option of forcing estimates at certain levels of aggregation to equal their true values before the addition of noise. Interior table cells were then raked to these aggregate cells. In this paper we discuss the advantages and disadvantages of adding noise to microdata as compared to cell suppression, and we describe the results of using noise and raking with the Research and Development survey.

KEYWORDS: Confidentiality; Disclosure; Noise; Cell Suppression

1. INTRODUCTION

The responding unit in many economic surveys and censuses conducted by the Census Bureau is the establishment. Individual establishments' responses are weighted (where appropriate) and aggregated, and estimates are generally produced by categorical variables like Standard Industrial Classification (SIC) code or geography. 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 is well known to data users.

The Census Bureau collects information from respondents under Title 13, U.S. Code, which prohibits the Census Bureau from releasing "any publication whereby the data furnished by a particular establishment or individual under this title can be identified." The disclosure limitation problem is to prevent data users from being able to recover any respondent's reported values using values appearing in the published tables. The Census Bureau 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).

2. CELL SUPPRESSION

The Census Bureau's current disclosure limitation technique used for establishment tabular data is cell suppression. Cells that pose a disclosure risk are 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 primary suppressions, or sensitive cells.

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 a 'D' flag.

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 very 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, the process must be performed separately for each data product. Among other things, this means that analysts 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 can't be used as complementary suppressions).

Coordinating suppression patterns among tables becomes impracticable in the presence of multiple requests for special tabulations. To truly prevent any disclosures, we would have to 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 we would have to keep an ongoing record of all interrelationships between all cells across all publication tables and special tabulations, a programming nightmare. We 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 for 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 D's. Data users frequently complain that we suppress too much data.

3. INTRODUCTION OF NOISE TO MICRODATA PRIOR TO TABULATION

3.1 General Description

An alternative method of protecting individual respondents is to add noise to their data. Suppose we perturb each respondent establishment's data by a small amount, say 10%. 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. By adding noise, we would avoid disclosing the dominant establishment's true value.

To each establishment in our sampling universe we would assign a multiplier, or noise factor. Whenever an establishment was canvassed in any survey or census, all of its values would be multiplied by the establishment's assigned noise factor. Within a particular survey or census, all establishments would have their values multiplied by their corresponding noise factors before the data were tabulated. Note that because the same multiplier would be used with an establishment wherever that establishment was tabulated, values would be consistent from one table to another. That is, if the same cell appeared on more than one table, it would have the same value on all tables.

Note that we would be adding 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 would rely on the assignment of the multipliers to control the effects of the noise on different types of cells. If the noise is added in a careful, systematic way (see Section 3.3), we can minimize its effect on cells that would not be suppressed under the usual procedure. Thus we can protect individual establishments without compromising the quality of our estimates.

3.2 The Multipliers

To perturb an establishment's data by about 10%, we would multiply its data by a number that was close to either 1.1 or 0.9. We could use any of several types of distributions from which to choose our multipliers. For instance, to perturb an establishment's data in a positive direction, we could choose the multipliers from a normal distribution with mean 1.1 and very small variance, perhaps .05 or .01. In any case, we would want to use a distribution centered at or near 1.1 and with small variance. If we wanted to ensure that all multipliers were at least 1.1, guaranteeing at least 10% protection to single-establishment cells, we could 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 avoidance 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 (but not exactly) 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 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.

Note that different establishments belonging to the same company should not have *exactly* the same multiplier. In some cases, different establishments owned by the same company compete with each other. Suppose two establishments from the same company each appeared in cells by themselves. One of the establishments could use its own true value and the value in the cell in which it appeared alone to derive its own noise multiplier. Then, if the establishment assumed that the same multiplier was used with all establishments belonging to its parent company, it could derive the other establishment's true value by using the known multiplier and the value in the cell in which the other establishment appeared alone. Thus in order to protect establishments from each other within the same company, each establishment in the company should have a slightly different multiplier.

3.3 Assignment of Multipliers and Its Effect on Estimates

We would like to assign the multipliers in such a way that we minimize the effect of the noise on those cells that are not at risk for disclosure. In particular, estimates at higher levels of aggregation are not generally sensitive. In economic censuses and surveys, the most common sub-national estimates are produced by SIC and geography. In assigning multipliers to establishments, we would like to arrange for these estimates to contain as little noise as possible.

One way to accomplish this is to ensure that among all establishments contributing to one of these estimates, the number having a positive amount of noise (multiplier > 1) and the number having a negative amount (multiplier < 1) are roughly equal. More precisely, considering that different establishments have different measures of size, we would like to ensure that the *absolute amount* of noise added and the *absolute amount* of noise subtracted roughly cancel each other out when summed over all establishments contributing to the estimate.

To this end, we would assign the multipliers in a systematic way. Before assigning multipliers, we would sort companies by SIC × geography × measure of size. Then multipliers would be assigned in a pairwise-alternating fashion, with the first establishment being perturbed in one direction and henceforth each successive pair of establishments being perturbed in the opposite direction from the pair immediately preceding it. The direction of the multiplier (greater than 1 vs. less than 1) for the first establishment would be chosen randomly.

To illustrate, suppose the first company was assigned a multiplier close to 1.1 (or, more precisely, all establishments belonging to the first company were assigned multipliers close to 1.1). Then the second and third companies would be assigned multipliers close to 0.9; the fourth and fifth companies, close to 1.1; the sixth and seventh, close to 0.9; etc. This procedure is "better" than assigning noise randomly because it assures that for every establishment that is assigned a noise factor > 1, there is in general another establishment of about the same size in the same SIC and the same geographical area that is assigned a factor < 1. Thus when aggregate estimates are computed, the noise present in these two establishments should have a tendency to cancel out.

Intuitively, both random assignment and systematic assignment of noise factors should provide that the expected value of the amount of noise present in any estimate is zero, thanks to the symmetry of the distribution of the multipliers. (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 estimate is simply the sum of the noise in its component establishments.) However, the systematic procedure bears in mind the skewed distribution of the measures of size among the establishments and should help to reduce the *variance* of the amount of noise as compared to random assignment. Aggregate estimates computed by SIC × geography should thus contain very little noise.

For other non-sensitive cells (aggregate estimates not computed along SIC × geography lines, detailed cells with many contributors, and detailed cells having few contributors but of roughly the same size), we would still have the result that, on average, estimates would not be altered by much. For these cells, it is still true that the establishments that are perturbed in the positive direction and those that are perturbed in the negative direction will generally balance each other out, although the sorting of the establishments doesn't ensure this as effectively as for the SIC × geography estimates. Here, the systematic assignment of noise factors would be similar to random assignment in terms of the average amount of noise present.

In contrast, a cell that is dominated by a single contributor would 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 estimate. Looked at another way, the more dominant the largest contributor, the more the amount of noise present in the cell estimate will resemble the amount of noise present in the largest contributor (about 10%). Thus the cells that would have been at greatest risk for disclosure would in general receive the most noise, and the noise in the cell total would prevent users from being able to recover an individual respondent's true value from the published value.

3.4 Adding Noise to Sample Data

In sample surveys, each respondent's data is generally weighted inversely proportionally 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. For sample data, to reflect the protection already provided by the sample weight, noise would be applied as follows:

For each establishment in a cell, calculate

$$\text{establishment value} \times \left[\text{multiplier} + (\text{weight} - 1)\right]$$

and then add up these noise-added establishment values 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.

This procedure has the effect of changing (weighted) values for certainty or near-certainty establishments (those having weights close to or equal to 1) by a large amount while changing weighted values 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 aren't protected by their weights. 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.

3.5 Updating Multipliers

Many surveys produce trend statistics, statistics that describe a percent change in the level of a variable from one time period to another. If we were to use the same multiplier for the same establishment in successive iterations of a periodic survey, we would be showing exact percent changes for establishments in single-establishment cells, and this would be a disclosure. (The common multiplier factors out of all level estimates, yielding the true percent change.) Multipliers would have to be changed from one period to the next in order to protect these trend statistics.

In order to preserve the utility of trend statistics, we would update multipliers in such a way that a particular establishment's values were always perturbed in the same direction in successive iterations of the survey. In other words, if the original (first) multiplier was chosen from a distribution centered at or near 0.9, we would choose all new multipliers from the same distribution. If the original multiplier was close to 1.1, we would choose all new multipliers close to 1.1. This way, if a particular estimate ends up being biased because of the addition of noise, in spite of our efforts to the contrary (see Section 3.3), it will at least be biased by roughly the same amount from one period to the next, thereby preserving the trend. Otherwise, if the direction of the bias were able to change from one period to another, the underlying trend might be obscured by the noise.

By using similar multipliers with a single establishment from one period to the next, we would maintain the longitudinal qualities of the data. And by varying the noise factors slightly between periods, we would avoid disclosing the exact value of any establishment's true percent change.

3.6 Using Different Multipliers for Different Data Items

In addition to protecting the values of an establishment's individual data items, we also need to protect the relationships between data items. For example, if in some survey an establishment reports its total revenue as well as components of the total like advertising revenue, we would need to protect the ratio of advertising revenue to total revenue for that establishment. This would only be a concern in single-establishment cells; as long as there were two or more establishments contributing to a cell, it would be impossible for any data user to distinguish any one establishment's exact share of each data item. In a single-establishment cell, however, it is known that 100% of each data item (with or without noise) is attributable to the sole contributor. Thus if the data items in both the numerator and the denominator of the ratio were multiplied by the same noise factor, then in computing the ratio the noise factor would cancel out of both the numerator and denominator, yielding the true ratio for the establishment.

To protect inter-variable relationships, we could use different multipliers for different data items. A base multiplier would be maintained for each establishment, and a different adjustment factor would be assigned for each item published. For example, say establishment A's base multiplier is 1.12 and establishment B's is 0.87. When tabulating total revenue, we might multiply establishment A's value by 1.123 and establishment B's value by 0.867. An adjustment of 0.003 has thus been added to (or subtracted from) the base multipliers for purposes of tabulating total revenue. When tabulating advertising revenue, we might multiply establishment A's value by 1.125 and establishment B's value by 0.865, in which case an adjustment of 0.005 has been added to (or subtracted from) the base multipliers. Thus when computing the ratio of advertising revenue to total revenue for either establishment, the different adjustments to the base multiplier for different data items would prevent exact disclosure of the ratio.

A major drawback to using different multipliers for different data items is that we could no longer guarantee that detail data items added to their proper totals within an establishment. One possible solution would be to define one selected detail data item as the difference between the aggregate data item and the sum of all other detail items. This would guarantee additivity but would have an unpredictable effect on the data item selected to be defined as the difference.

3.7 Flagging Cells with a Large Amount of Noise

The percentage of noise in a cell would be defined as the percent by which the noise-added value for the cell differed from the true noise-free value. Thus we would have to calculate both the noise-added and noise-free values for each cell in order to quantify the amount of noise each cell contained. Noise-added values would be used in all data released to the public, while noise-free values would be used for internal research and analysis.

All resulting table cells containing a large percentage of noise, say a 7% change in value or more, would be flagged so users would know that the values may not be useful. This set of cells would 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 would explain how and why noise was added and would let users know that disclosure limitation had been performed. Users would thus be discouraged from using the inaccurate, noise-added totals in sensitive cells to try to recover a value for any individual contributor.

We could 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 would at least *think* the cell contained a lot of noise and would hesitate to treat the cell value as reliable. We would expect relatively few cells of this type.

Note that we aren't concerned with whether a cell appears to fail the disclosure rule *after* noise is added. The fact that a single establishment's noise-added value appears to dominate the overall noise-added cell estimate does not imply that a data user would be able to determine the establishment's true noise-free value. Users depend on an accurate cell total to be able to recover the value for any individual contributor to the cell, and we also don't need to protect noise-added values.

Cells exceeding the noise threshold, as well as sensitive cells not sufficiently protected by the noise, would 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 didn't 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. BENEFITS OF NOISE

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. Each establishment would need only to have its data items multiplied by the establishment's noise factor, possibly with different adjustments to the noise factor for different data items, prior to tabulation. In each table, we would still have to identify those insufficiently-protected sensitive cells (cells that would normally be primary suppressions) in order to flag them, but the complicated and lengthy process of choosing complementary suppressions would be avoided. The addition of noise would not be table-specific, whereas complementary suppressions must be identified on a table-by-table basis. Computer programs for adding noise would also be 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 would eliminate the need to coordinate cell suppressions between tables. Under the current cell suppression practices, disclosure analysis must be done separately for each data product. This 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. (Otherwise, for instance, a cell that is used as a complementary suppression on one table might appear unsuppressed on another table.) Keeping track of suppressions is difficult to orchestrate and difficult to understand. However, using noise to protect estimates would make this unnecessary. For each data release, we would need only to identify and flag any cells that were primary suppressions or that contained more than the prescribed acceptable level of noise.

In particular, eliminating the need to coordinate suppressions would allow for easy and quick fulfillment of requests for special tabulations. Using noise to protect estimates would allow us to compute as many special tabulations as we needed without having to keep track of what estimates had already been released. We would again need only to identify and flag any primary suppressions and cells containing too much noise.

The addition of noise was designed mainly to overcome the multiple special tabulation problems that arise with cell suppression, but it also may allow for more valuable data to be published in our standard releases. With cell suppression, users lose information both for cells which are primary suppressions and for those that are complementary suppressions. With the noise technique, because of the systematic way in which noise is added to the establishments (see Section 3.3), 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 some that would have been complementary suppressions. Thus for publications which normally contain many complementary suppressions, the noise technique should provide data users with more valuable information.

Note that adding noise would not help much with tables where most suppressions are primary suppressions. For those tables, only a reduction in detail could reduce the number of cells for which data was suppressed or severely altered.

5. ARGUMENTS AGAINST NOISE

5.1 Insufficient Protection for Single-Establishment Cells

Some respondents may not feel that the added noise provides enough protection to values in single-establishment cells. Under the cell 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.

On the other hand, for every value that is suppressed under the cell suppression approach, an interval which contains that value can be derived. For example, if a value of 100 is suppressed, users can look at surrounding cells to determine that the value is between, say, 84 and 124. Users often derive this interval and then use the midpoint as an estimate for the cell value, in this case 104. Sometimes the midpoint is very close to the true value, and other times it is not.

Which method offers better protection, noise or cell suppression? This is a subjective question and is not easily answered. In fact, the answer may ultimately depend on the opinions of our respondents. The point is that the question of whether noise provides *enough* protection to sensitive cells could just as easily be directed at the cell suppression approach.

5.2 Perceptions of Data Quality

It is possible that some respondents may resent putting time into preparing good responses if they know the Census Bureau is only going to add noise to them anyway. We would need to emphasize that we were not simply adding noise indiscriminately. Noise would be added in an unbiased, controlled way so as to preserve the statistical properties of the data while having a negligible effect on non-sensitive estimates. To maintain the properties of the data, we would need to know what those properties were to begin with. And to assess the effect of the noise on important aggregate estimates, we would have to know their true noise-free values as accurately as possible. Thus it is crucial that we begin the noise addition process with the true values in order to perturb the data in a predictable way.

Another important consideration is that the Census Bureau makes extensive internal use of respondents' data for such tasks as imputation, sample redesign, and regression analysis. To perform these functions accurately, we need as our inputs the best estimates that respondents can provide.

There may be concern on the part of some data users as to the quality of the data after noise has been introduced. In using this proposed technique, we would hope that the users' desire for multiple special tabulations and their desire to see more published cells (at the expense of noise) would outweigh their desire for true values (at the expense of suppressions). Flags would inform users of the data's utility by drawing their attention to cells that had been adversely affected by the noise. Also, users know that our published estimates for surveys already have sampling error associated with them, as described by the CV, and are not true, exact values. Even our census values contain some "noise" due to various types of nonsampling errors (reporting errors, keying errors, imputation, etc.). In general, users know that we are publishing our best possible estimate of each cell's value.

The fact that we would be deliberately perturbing the data, however, may lead users to feel that the numbers they see in the tables are not our best possible estimates. While the other types of error that are already present in our published values are errors that we attempt to control or eliminate, the added noise would be error that we were *actively introducing* into our estimates. We would have to emphasize the purpose of adding noise and remind users that it was added in a way that would minimize its effect on non-sensitive estimates.

It should be mentioned that some surveys may not warrant the use of the noise technique at all. Many surveys publish estimates at such high levels of aggregation that few, if any, cells are suppressed under the current procedure. If a survey publishes very few sensitive cells and doesn't entertain requests for detailed special tabulations, there would be no need to compromise data quality by introducing noise into the estimates because there is very little, if any, information that requires protection. The noise technique would be useful primarily with surveys that receive many requests for special tabulations or that currently suppress large numbers of cells.

6. VARIATIONS ON ADDING NOISE

6.1 Noise with Some Cell Suppression

As mentioned in Section 5.1, many people feel uncomfortable about the idea of publishing any value, even one with a lot of noise in it, for cells with only 1 or 2 establishments. To address this issue, they have asked if noise could be applied *and* those particular cells could still be suppressed, along with a sufficient number of complementary suppressions. (Other cells that were identified as sensitive but that had 3 or more establishments in them would not be suppressed and would be protected by the noise and the accompanying flag.)

We could suppress one- and two-establishment cells, but there are several disadvantages to this approach. Two procedures would have to be applied to the data, thus making disclosure limitation more time-consuming and harder to understand. There would still need to be suppression pattern coordination among all tables, although there would be fewer suppressions to coordinate. In particular, this would not solve our problem with multiple special tabulation requests. This approach seems to possess the disadvantages of both cell suppression and noise, and it does not solve our problem with multiple special tabulations.

The fact that only a flag, and *not* a value, would appear in the cell may help reinforce the idea that the cell's value is protected by the noise. As mentioned earlier, however, the decision as to whether the noise and the flag offer *enough* protection may ultimately rest with the respondents.

6.2 Cell Suppression for Standard Releases and Noise for Special Tabulations

Some people have asked about the possibility of using cell suppression for all standard publications and using noise for all special tabulations. This practice could compromise the protection provided by cell suppression. A special tabulation could contain some of the same cells that appeared in a standard publication. For cells that were primary suppressions in the standard release, this would not be a problem because in the special tabulation these cells would contain a lot of noise and would be flagged. However, cells that were suppressed purely as complementary suppressions in the standard release should not receive much noise in the special tabulation and thus would not be flagged. A user could substitute the values from the special tabulation, which would be relatively noise-free, into the corresponding suppressed cells in the standard table and through addition and subtraction could obtain close approximations of some primary suppressions. Thus we would lose much of our protection for the primary suppressions.

Another problem with using noise only in special tabulations is inconsistency between tables. If a special tabulation contained some of the same cells that appeared in a standard publication, these cells would contain noise in the special tabulation but not in the publication. It would be inconsistent to have the same cell appearing in two locations with a different value in each location.

6.3 Adding Noise and Raking to True Values

6.3.1 General Strategy

One of the main objections to the idea of adding noise is that it would affect *all* estimates, not just those that would have been disclosure risks. An alternative that would address this problem is to add noise but then force the published values of the more important estimates (presumably those at higher levels of aggregation) to equal their true (without noise) values. Interior cells would then be raked, or proportionally adjusted, so that they still summed to the aggregate estimates. ("Raking" is also known as iterative proportional fitting.) Fagan and Greenberg (1985) provides some background on raking.

We would first need to determine the level of aggregation at and above which we wanted published estimates to equal their true values, i.e., their values uncorrupted by noise. This should be a level at which we expect very few, if any, sensitive cells. We would proceed to introduce noise into all establishments and compute all estimates with noise present. Then, at the lowest level of aggregation at which published estimates were to equal their true values, we would force the estimates with noise present to equal the corresponding values with no noise, raking the interior cells as well to preserve additivity and proportions. For multi-dimensional tables, the interior cells would have to be simultaneously raked to all marginal totals that were held fixed at their true values. All estimates at levels of aggregation higher than the level at which we raked would automatically equal their true values, since they would be summations of components that had already been raked to their true values.

This method would leave many non-sensitive cells, namely those in which true values would be published, totally unaffected by noise. At the same time, the raking should not have a very pronounced effect on any of the other estimates. Since estimates at higher levels of aggregation, where the raking would be done, shouldn't have much noise in them to begin with, the raking factors would be relatively small. Those cells that had a lot of noise before raking would still have a lot of noise after raking, so the noise would still offer protection to sensitive cells.

There are two ways to approach the raking, as described in the next two sections. One option is to rake each table individually; the other is to rake to all fixed marginal totals simultaneously before producing any tables. Each has advantages and disadvantages as compared to the other.

6.3.2 Raking Each Table Separately

One approach to raking is to rake each data product individually. This has the advantage of allowing analysts to determine on a table-by-table basis (including special tabulations) the lowest level of aggregation at which the estimates will not be disclosure risks. The interior cells in a particular table would then be raked only to the selected marginal totals that actually appeared in the table, regardless of what other marginal totals might be held fixed in other tables. This would in general allow more cells to equal their true values than in the case where estimates are simultaneously raked to all fixed marginals. This is described in Section 6.3.3, and is related to the level of detail in the table.

An establishment would retain the same *base* noise factor throughout all tables. However, since each table would be raked individually, the *net* amount of noise (after raking) present in an establishment's contributions to different cells could differ from one table to another, and even from one column (or row) to another within the same table. This could create problems of consistency between tables. One requirement to maintain consistency is that if a group of cells are published with no noise added in one table, they should be noise-free in all tables in which they appear. Otherwise the same cell could have different values in different tables. We would therefore have to keep track, from one table to the next, of what cells had previously been held fixed. This includes keeping track of noise-free cells when producing special tabulations. This process would be similar in complexity to keeping track of suppression patterns among tables under the cell suppression scheme. If we were to use this technique, it would probably be best reserved for small-scale surveys having tables that were relatively non-interrelated.

6.3.3 Raking Once Before Tabulating

One way to avoid having the same cell appear in two different tables with two different values is to rake all estimates only once. Before producing any tables, analysts would determine the set of cells that would be forced to equal their true values. Then all other cells from all publication tables would be simultaneously raked to all of these fixed cells.

This single raking would be done by first constructing an $n \times n$ "supermatrix," where n is the total number of categorical variables that appear in any of the tables. After adding noise to all establishments, each establishment would then be tabbed into *exactly* one interior cell of this supermatrix, depending on its values of the n categorical variables. Then, an n_0-dimensional raking would be done, where n_0 is the number of categorical variables for which some or all of the marginal totals were to be fixed at their true values ($n_0 \leq n$). The raking would define an adjustment factor for each cell, indicating the percentage by which the raking changed the value in the cell. This raking factor would be applied to each establishment contributing to the cell. The net noise factor for the establishment would then be the product of the original noise factor and the adjustment factor determined by the raking.

This net noise factor would then be associated with the establishment throughout the production of all standard tables and special tabulations. This would guarantee consistency of estimates between tables because the same set of establishments in a cell with the same set of noise factors would always produce the same estimate. And producing special tabulations would require no special procedures; we would simply tabulate each establishment's value, multiplied by the single factor that reflects both noise and raking, and then flag appropriate cells. We would not have to worry about keeping track of what estimates had appeared in previous tables and how much noise they contained.

The main disadvantage of the single-rake approach is that it would limit the level of detail at which we could force estimates to equal their true values. In the presence of a large number n of categorical variables, interior cells in the supermatrix would generally be sparsely populated. If we tried to rake the interior cells to too many fixed marginal totals simultaneously, some cells may then be constrained to equal their true values to guarantee (sometimes trivial) additivity in all dimensions of the supermatrix, thereby cancelling the effect of the noise. Analysts would have to limit the number of marginal totals that were held fixed; otherwise, the raking could undo whatever protection was provided to interior cells by the addition of noise, which could leave many sensitive cells unprotected.

7. RESULTS WITH ACTUAL SURVEY DATA

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

To assign multipliers, we first sorted the sample companies by SIC grouping × R&D expenses (descending). We then assigned multipliers to the companies in the pairwise-alternating fashion described in Section 3.3. Note that if we had strictly followed the strategy described in Section 3.3, we would have assigned multipliers to the sampling *frame* before actually selecting the sample. However, in this case we were working with a survey that had already been conducted, and we no longer had access to the frame. Also, the main reason for assigning multipliers at the frame stage is that the same frame can be used for many surveys, thereby avoiding duplication of effort in assigning multipliers. For testing purposes a multiple-use frame was not an issue, so assigning multipliers to the sample file seemed the most efficient way to proceed.

To generate 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 std. dev. $\sigma = .02$: N(0.9, 0.02) and 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 --- $X \sim .1\ B(6,2) + 0.8$ and $X \sim .1\ B(2,6) + 1.1$

After generating a multiplier for an establishment, we applied the multiplier to the establishment's data items using the formula in Section 3.4, since we were working with sample data. To maintain simplicity, we didn't use different multipliers for different data items as described in Section 3.6.

We reproduced Table 2 from the R&D publication, which shows R&D expenses broken out by federally-sponsored vs. corporate-sponsored, for the 26 SIC groupings. (Table 7.1 below shows the structure of R&D Table 2.) We ran 100 simulations of Table 2 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 percents 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 greater 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 1000 replications of Table 2, using the Beta distribution to generate multipliers, and computed summary statistics to describe the behavior of the cells over all replications. To isolate the effects of the noise, we did not perform any raking at this stage. Below is a copy of Table 2 showing, for each cell, the ratio of a) the average of the 1000 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 simply numbered 1 through 26 and do not appear in the same order as in the R&D publication. Sensitive cells are shaded.

Table 7.1 Ratio of Noise-Added Value to Noise-Free Value

average noise-added value over 1000 simulations of R&D Table 2, divided by true unperturbed value

stub #	total R&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

Note that the values are indeed close to 1, for both sensitive and nonsensitive 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 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, given that the sample has already been selected.

It remains to be seen whether bias is introduced by noise over repeated applications of the entire sample selection and estimation process. Replicating the entire process of selecting a sample, adding noise to the sampled establishments, and tabulating was not feasible in this case because of the unavailability of the sampling frame file, the unavailability of actual values for all establishments in the frame, and the need to rerun the sampling and estimation programs repeatedly.

Note that while the expected value of the amount of noise in any one *establishment* is 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 the degenerate case where an estimate is composed of only one establishment, the estimate will 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 1000 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 $\hat{Y}_{noise}$ is simply the standard deviation of the noise itself: writing $\hat{Y}_{noise} = \hat{Y} + e$ and taking $s(\hat{Y})$ and $Cov(\hat{Y},e)$ to be zero, we have $s(\hat{Y}_{noise}) = s(e)$. The value in the table, $s(\hat{Y}_{noise})/\hat{Y}$, can be thought of as the coefficient of variation (CV) of the noise-added estimate, given the noise-free estimate, i.e., $CV(\hat{Y}_{noise}|\hat{Y})$. Table 7.2 below shows these "CVs" over the 1000 replications. Again, sensitive cells are shaded.

Table 7.2 "CVs" of Noise-Added Values

std. dev. of the 1000 simulations, divided by the true noise-free estimate

stub #	total R&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

std. dev. of the 1000 simulations, divided by the true noise-free estimate

stub #	total R&D	federal	company
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

Note that the CVs are generally much higher in the sensitive cells than in the nonsensitive ones. The variability of the amount of noise present in sensitive cells is much greater, so a sensitive cell should be much more likely than a nonsensitive 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 cell, we computed the *absolute* percent noise present in the cell for each replication. We then computed an overall "percent noise" by averaging these absolute percentages over all 1000 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 "percent noise" variable among cells of various types. Table 7.3 below gives the results.

Table 7.3 Amount of Noise in Cells, By Type of Cell

	amount of noise			
% noise in:	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

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 percent.

Cells that would have been primary suppressions receive noticeably more noise than nonsensitive 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 don't expect *all* sensitive cells to receive a lot of noise (see Section 3.7), nor do we expect that none of the nonsensitive cells will receive a lot of noise. Table 7.4 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 didn't receive much.

Table 7.4 Counts of Cells Having Large vs. Small Amounts of Noise

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

This table further illustrates that the noise technique generally leaves nonsensitive cells (including marginal totals) relatively noise-free, while most sensitive cells receive a lot of noise. The few sensitive cells that don't exceed the noise threshold would be flagged as described in Section 3.7, along with both sensitive and nonsensitive cells that do exceed it.

We also wanted to test the effect of raking after noise was introduced. The sample size in the R&D Survey is relatively small, and the amount of detail in any of the publication tables is limited. In fact, there were only three estimates that were at a sufficiently high level of aggregation that we felt they could safely be held fixed at their true values without creating any disclosure risk. These were the totals over all SICs for corporate R&D expenses, federally-funded R&D expenses, and total R&D expenses. With only three marginal totals to rake to, we chose the single-raking option described in Section 6.3.3. To differentiate between the effects of the noise and the raking, we analyzed the results of adding noise both before and after raking.

We raked each of our 1000 replications individually and recomputed the "percent noise" variable for each cell by averaging the absolute percent noise for the raked cell values over all replications. Table 7.5 shows the effect of the raking by showing how many cells contained more vs. less than 7% noise before and after raking.

Table 7.5 Cells Having Large vs. Small Amounts of Noise, Before vs. After Raking

			after raking		
			|noise| < 7%	|noise| ≥ 7%	
before raking	|noise| < 7%	sensitive	0	1	1
		nonsensitive	57	2	59
	|noise| ≥ 7%	sensitive	0	10	10
		nonsensitive	4	3	7
all		sensitive	0	11	11
		nonsensitive	61	5	66

Notice the upper right and lower left boxes in the main part of the table. In the upper right corner, we see that only 3 cells (2 nonsensitive, 1 sensitive) were below the noise threshold (7%) before raking but exceeded it afterward. And in the lower left, only 4 cells (all nonsensitive) started out with more than 7% noise but ended up with less after raking.

Raking factors are determined by how much noise is present in the marginal cells whose values are being held fixed. Since the assignment of multipliers assures that there is no tendency for the percentage of noise in a marginal cell to be positive rather than negative, the raking factor is equally likely to be greater than or less than 1. Thus we would expect that for every cell that started out with more than 7% noise (in absolute terms) before raking but ended up with less than 7%, there would on average be another cell whose noise percentage changed in the opposite direction with respect to the 7% threshold. Among our limited number of cells, this appears to be roughly the case.

Also, we would expect relatively few cells to cross from one side of the threshold to the other as a result of raking (only 7 out of 77 in our example). These would be cells that contained close to 7% noise before raking, so that a small percent change in the cell value could push it over the threshold. For the most part, cells that started out with a lot of noise still had a lot after raking, and those that started off below the threshold remained below it. In other words, the raking procedure doesn't counteract the protection provided by the noise. This is not surprising, since we don't expect percent changes due to raking to be very large. Raking factors are determined by the amount of noise present in the marginal cells that are to be held fixed, and these marginal cells should end up with little noise because of the way the multipliers were assigned.

8. CONCLUSIONS

Adding noise to establishment microdata has clear advantages over cell suppression as a way of providing the required protection to individual respondents. As we move into an era of customized data products and user-defined tables, the noise technique would afford us the flexibility to accommodate a wide variety of data requests without the worry of inadvertently disclosing any particular respondent's values. Unlike with cell suppression, we wouldn't have to keep track of all prior requests in order to guarantee that each new data product was free of disclosures.

The results of our test with the R&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 nonsensitive cells; thus noise provides protection where it is most needed.

Looking beyond the behavior of level estimates, we plan 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. We also plan to test the noise technique with some special tabulation requests. In this way we will see if we can use noise to protect respondents' data while ensuring that the data meet the varied needs of the users.

As mentioned earlier, the noise technique is probably not suitable for all Census Bureau 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 microdata could become a viable alternative to cell suppression for disclosure avoidance with establishment tabular data.

REFERENCES

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.

Evans, B.T., and Zayatz, L. Using noise for Disclosure Limitation of Establishment Tabular Data. *Statistical Research Division Report Series*, Bureau of the Census, to appear in 1996.

Fagan, J. and Greenberg, B. (1985). Algorithms for Making Tables Additive: Raking, Maximum Likelihood, and Minimum Chi-Square. *Statistical Research Division Report Series*, Census/SRD/RR-85/12, Bureau of the Census.

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