Recent Developments in the Use of Noise for Protecting Magnitude Data Tables: Balancing to Improve Data Quality and Rounding that Preserves Protection

Paul B. Massell
U.S. Census Bureau
Paul.B.Massell@Census.Gov

Jeremy M. Funk
U.S. Census Bureau
Jeremy.M.Funk@Census.Gov

Keywords: disclosure avoidance, magnitude data tables, microdata, statistical noise

Abstract

Over the years, many program managers in the Economic Directorate of the U.S. Census Bureau have been interested in exploring alternatives to cell suppression for protecting released tables, i.e., protecting the confidentiality of microdata that underlie the tables. The goal is to produce tables with fewer suppressions that are still fully protected. With fewer suppressions, the tables would likely be of greater value to users, even if the cell values were perturbed a bit from their original values. A decade ago, a method was developed by researchers at the Census Bureau Evans-Zayatz-Slanta (J. Official Statistics, 1998) that involves adding noise to the microdata. Recently this method has been applied to survey and census tables with distinctive features. In some programs there may be a table that is considered the primary table, and one may wish to fine-tune the noise method, using a technique called 'noise balancing', so that the data quality is as high as possible for this table while maintaining the quality in other tables. There are also differences in the type of rounding that is applied to the raw data as it is transformed into microdata and perhaps later when microdata values are summed to form cell values. Rounding often adds uncertainty about the pre-rounded value, so it would seem to increase disclosure protection. However, when noisy values are rounded, this sequence has the potential of reducing the protection level of the noise. To ensure that this does not happen requires enhancement of some of the ordinary rounding methods.

1 Background: Protecting Magnitude Data Tables

There are now several methods for protecting magnitude data tables. Since cell suppression has been used extensively at the Census Bureau for many years, it makes sense to describe what aspects of cell suppression are seen as drawbacks. We describe below how Evans-Zayatz-Slanta (EZS) noise overcomes most of these drawbacks. Some of these drawbacks we believe apply to controlled tabular adjustment (CTA) but since our experience with CTA is limited, and because CTA is still undergoing rapid development, we will restrict our comparison discussion to EZS noise versus cell suppression (Dula, Fagan, Massell, [2]).

Tables protected by cell suppression typically provide no explicit information about the suppressed cells; this applies whether such cells are sensitive (i.e. primary suppressions) or are non-sensitive cells simply used to protect sensitive cells (i.e., secondary suppressions). In theory, a user with linear programming skills could write an audit program to compute an uncertainty interval for each suppressed cell. He would then have a range for the suppressed cell value, but the interval, if wide, would not be useful for estimating the true value of any individual respondent's contribution. We claim that such wide intervals are common in practice. More realistically, most users don't have the time or inclination to write such programs. They would rather be directly supplied with an approximate value for each cell, as long as they were assured that the approximated values were reasonably close to the actual values. EZS noise is one way of providing such useful approximate values.

We use the expression 'deterministic protection methods' to refer to a set of protection methods that are generally more mathematical rather than statistical in nature. This includes cell suppression and controlled tabular adjustment. These methods typically do not use random number generators. The search aspect of the method is performed using a mathematical algorithm, for example if linear programming is used then some version of the simplex method is typically utilized. A more general definition is that deterministic protection methods are methods in which the office determines the size of an uncertainty interval that should be constructed about a given sensitive value. The office then uses an algorithm that guarantees that the perturbations or suppressions will produce an uncertainty interval around the value at least that wide. This can be done because both the determination of required uncertainty and the creation of uncertainty are being performed at the cell level.

2 Overview of EZS Noise

In the EZS noise method the required uncertainty is still determined at the cell level, but in contrast to cell suppression it is not created there. Under this method the values of both sensitive and safe cells are perturbed indirectly via the underlying microdata contributions to the total cell value. The difference between these two classes of methods is perhaps best seen through an example. Suppose in a deterministic method such as cell suppression, an office indirectly states that the true value of some cell lies in the interval [90,110]. In a noise method, one might release the value 105 and simply state that it represents the result of selecting a noise multiplier for each contribution to the cell from a reasonable noise distribution. In some cases, the office might decide to reveal some information about the noise distribution or even reveal it completely. Even when revealed completely, however, the user does not know enough about any given cell value to estimate a single contribution accurately (Massell, [6]).

An important general issue is the need to protect all tables generated from a given microdata set in a consistent way. When the protection method is cell suppression, this involves a technique called 'back-tracking'. This process involves ensuring that if a given suppressed cell is supporting x units of 'protection flow' in one table, that it supports at least x units in all other tables in which it appears. In other words, we need to ensure that no table allows a good estimate of any suppressed cell; if it did, one could use that good estimate to begin unraveling the suppression pattern.

Another issue that is less general is the Census Bureau requirement to protect economic data at the company level. This can get complicated because many companies, including a high percentage of the largest ones, have multiple establishments in several locations. For these multi-unit companies, the microdata records typically contain data for only a single establishment, or some group of establishments, but not the whole company. Protecting at the company level entails protection of the sum of all the establishment values as well as the values separately. Meeting the company level protection requirement with cell suppression requires complex code (e.g., computing the 'capacity' of every table cell to protect a given sensitive cell). Company level protection is much easier to implement when using EZS Noise.

2.1 Properties of EZS Noise that Overcome Drawbacks of Cell Suppression

EZS noise allows approximate values to be released for all cells. In some cases the office may decide that to prevent even the appearance of a disclosure risk it is best to suppress all the sensitive cells, or at least those based on a very small number of contributions. In any case, EZS allows the publication of those values that would be secondary suppressions under cell suppression.

With the EZS method, since noise is added at the microdata level, all tables generated from the same perturbed microdata are protected consistently. In other words if a cell value appears in two or more different tables it will have the same (noisy) value in each table. This property also holds for any variant of the basic EZS method, since each record is assigned a single permanent noise factor.

Protecting at the company level is easy to implement with EZS Noise. This can be accomplished simply by assigning noise in the same direction (+/-) to all microdata records associated with establishments from the same company. For example, if the minimum noise magnitude is 10%, this direction rule requires that all establishments (for a given company) are assigned noise factors that are greater than 1.1, or all are assigned noise factors that are less than 0.9. This rule ensures that the sum of these noisy establishment values, i.e. the noisy company value, will also be perturbed by at least 10%, either up or down from its true value.

Another advantage of EZS is that since the noise assignment occurs at the microdata level the protection of respondent values are local; it does not depend on the structure of the table in which the contributing cell exists. In particular, tables of high dimension and/or tables with hierarchies can be protected as easily as a simple two dimensional table. This is particularly beneficial to the Census Bureau due to the size of our datasets and the complex inter-relationships of tables within and between survey programs. Coordination between several table releases based on the same microdata is automatic, unlike cell-suppression where intense efforts must be made to ensure that suppression patterns in different tables do not reveal information that is protected in another table.

The basic version of the EZS method can be implemented by a good statistical programmer. The code may only be a couple of pages in any of the common statistical packages. An office can use a simple rule, such as the P% rule, for determining which cells are sensitive and how much perturbation would be required from noise if it were the only source of uncertainty. For EZS noise, the office needs to use a noise distribution that is calibrated to the sensitivity rule and uncertainty measure being used. For this calibration, it is useful to consider the effect of the noise on the most sensitive cells such as those with only 1 or 2 contributors.

To test whether the code is correct and that a reasonable noise distribution has been selected, it is useful for an office to generate analysis tables or graphs that summarize the behavior of the method for all sensitive cells and for all safe cells. The percent change of individual sensitive cells can be compared to the percent changes suggested by the P% rule for these cells. This will indicate whether the desired amount of perturbation is being applied to these sensitive cells. Looking at the percentage change in safe cells will illustrate the magnitude and range of perturbation to these cells and be a valuable measure of the effect noise has on data quality.

3 Measuring the Effectiveness of a Perturbative Protection Method

Consider any protection method for tables in which cell values are perturbed. This perturbation may be generated in a deterministic way (e.g., controlled tabular adjustment), in a stochastic way (e.g., the EZS noise method), or using a method which combines both deterministic and stochastic aspects. For any such method there is a desirable amount of perturbation, or a desirable range of perturbation that depends primarily on the sensitivity status of the cell. In our examples below we use the standard P% rule for determining both the sensitivity status of each cell and the desirable amount of perturbation. Simply put, the P% rule calculates an amount of suggested protection for each cell. If this value is less than 0 then the cell is declared 'safe' and no perturbation is required. On the other hand if this value is greater than 0 then the cell is declared 'sensitive' and the suggested amount of perturbation is equal to this value. In a deterministic method, it may be possible to meet these goals for all cells in some or all tables. In the EZS noise method (and perhaps more generally in all stochastic methods), these protection goals cannot always be met for all cells. That is, the statistical office (SO) has to be willing to tolerate some under-perturbation of sensitive cells and/or some over-perturbation of safe cells. For most reasonable noise distributions applied to most real microdata and tables, one would expect a little of both.

It would be nice if a SO could predict the amount of under and over-perturbation, or at least upper bounds for them, so that when the selected noise distribution is applied to a set of microdata the SO would know beforehand that the errors will likely be small enough to meet the SO's standards. Ideally this could be done through precise modeling of the data and of the effect of applying the EZS method to the data. In many situations this modeling approach would be very time-consuming, and is not realistic. A more realistic approach would be for the SO to perform enough simulations during the experimentation phase to gain confidence that the final simulation, performed as part of the production tabulation, would very likely produce errors that are under the SO's limits. This typically will require some trial and error analysis to find acceptable parameters for the noise distribution(s) the SO wishes to consider. We call this process calibration of the noise distribution.

3.1 Protection Multipliers: A Distributional Measure of Under-Perturbation

We have found distributional measures of under and over-perturbation to be more useful than the scalar measures mentioned above. The only disadvantage is that a density must be produced, either in the form of a table or graphically. To measure under-perturbation, one computes for each sensitive cell a protection multiplier, defined as the ratio of the absolute perturbation divided by the protection suggested by the P% rule,

$$\text{PM} = \frac{|\text{Perturbation from Noise}|}{\text{Suggested Protection}}$$

It is useful to compute the percentage of all sensitive cells that have PM ≥ 1. If that percentage is low then the amount of noise may need to be increased. Also of interest is the distribution of PM values less than 1. If a high percentage of these PM values are less than say 0.5, a SO may wish to add more noise. Alternatively, the SO may decide to suppress all sensitive cells with PM < 1, or those with PM < T, for some threshold T. Of course, such suppressed cells may still be easily recoverable unless the SO takes the computationally costly step of running a complementary suppression program and applies it to those cells.

3.2 The Percentage Change Distribution: A Global Measure of Noise's Effect on Data Quality

The distribution of percent changes to a designated subset of cells is an important indicator of the effect noise has on the quality of the published table. Ideally, a high percentage of all safe cells will be changed by small amounts, say between 0 and 3%. If a significant percentage have been changed by more than say 5%, the SO may try to lower the amount of noise being added by adjusting the noise distribution or may consider using the balancing version of the EZS method described below.

3.3 The Application of EZS Noise Factors for Un-Weighted Data

Let X = original microdata value
Let Y = perturbed microdata value
Let M = noise multiplier (or factor) drawn from a specified noise distribution

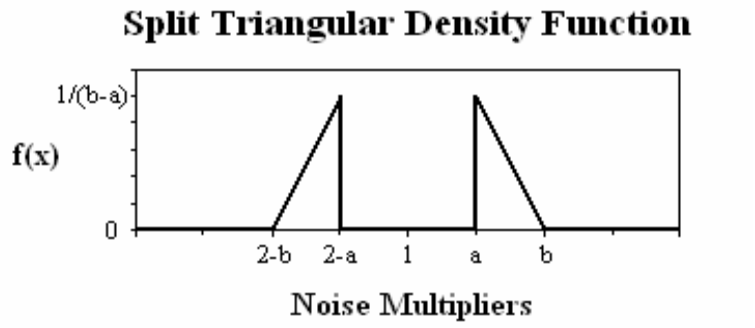
$$Y = X * M$$

The distribution we used for our examples below is the "split" triangular distribution, for which the density function is described below and illustrated in the figure below.

Let $1 < a < b < 2$

$2-b < x < 2-a$	$f(x) = k \bullet (x - (2-b))$
$a < x < b$	$f(x) = (-k) \bullet (x - b)$
otherwise	$f(x) = 0$

Here $k = (1/(b-a)^2)$ since the area under the density curve must equal 1.



The density is piecewise linear and is symmetric about 1. In our examples, we use a = 1.10 and b= 1.20.

4 Random Noise vs. Balanced Noise

A major criticism of using random noise for disclosure avoidance is that it may add excessive amounts of distortion to cells that would be shown much more precisely under deterministic methods such as cell suppression and controlled tabular

adjustment. A natural question to ask then is whether or not there is a way to modify the method such that it adds less noise to the non-sensitive cells, while retaining the amount of protection provided to the sensitive cells. One of the primary benefits of EZS noise over cell suppression is that it allows for the release of more usable data. Suppression has the advantage that all published cell values are the best estimates collected and produced by an agency. If noisy cell values are highly distorted then the benefit of noise over suppression is significantly reduced. As a result, we have investigated possible methods to reduce the overall amount of noise added to the data without compromising the level of protection.

One way to reduce the level of distortion to tables is to use a balancing algorithm that works to minimize noise for certain groups of records, in particular table cells. There are several issues to consider when using this type of method. Primarily the fact that respondents may contribute to multiple non-sensitive table cells within each of which it may be desirable to minimize the amount of noise applied. First of all, in hierarchical tables respondents usually contribute to cells at many levels of hierarchy. There may also be many establishment level records that belong to the same company but contribute to various cells throughout the table. The Census Bureau requirement of company level protection necessitates that respondents from the same company get noise in the same direction. The problem then becomes choosing which cells to use in assigning balanced noise factors and is discussed in detail in the following paragraphs.

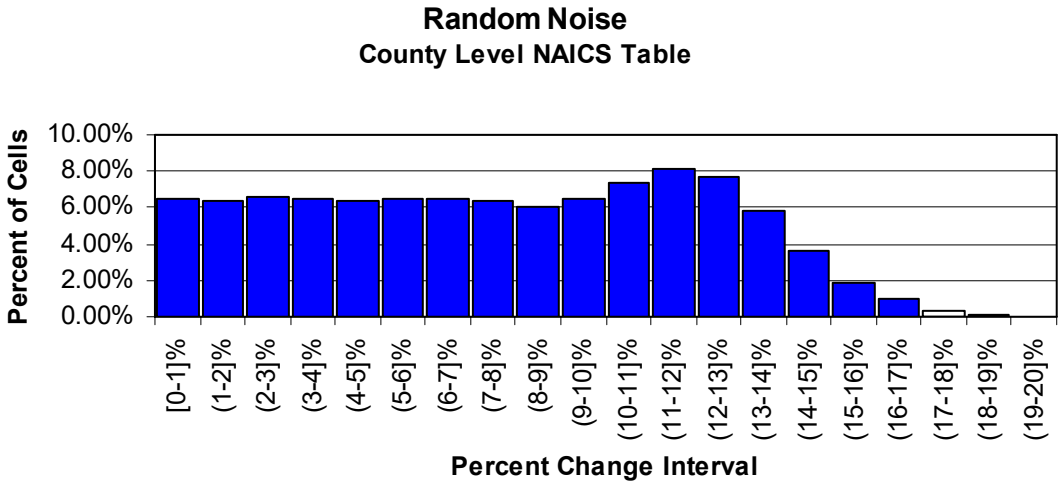
There may be ways to determine the optimum pattern of noise direction allocation for tables in which records contribute to multiple cells, but such a method would almost certainly increase the complexity of the EZS method significantly. Since ease of implementation is an important and attractive feature of EZS, we would like to preserve it. The solution we have come up with is to define a subgroup of table cells and apply noise balancing to these cells. To keep the method simple we need to be able to determine a list of cells such that each respondent contributes to exactly one cell. In this way we can sort the complete list of records according to this list of cells (we call this list the assignment sub-table), and assign noise factors one record at a time in a single pass through the data set. This allows us to keep the method implementation simple and computation undemanding.

The primary drawback to this method is that it only allows for noise to be balanced on a subset of all the cells that we would like to publish. We therefore have to consider its effects on all table cells not in the assignment sub-table, including aggregate cells and cells from different tables with different structures. In hierarchal tables we expect that balancing the noise in cells of a high level of detail will 'trickle up' into the aggregate cells. This is because summing cells with less noise will effectively produce aggregate cells with less noise. It also may be the case that if certain tables are slightly different although highly correlated, balancing the noise according to one may also have a positive effect on the other.

To present an example of how balancing can work to improve the quality of non-sensitive cells we will use sample data from the US Census Bureau's County Business Patterns Survey. The test data set we used consisted of all the employer business establishment level data from a single US state. The table of interest was County by NAICS (detailed industry at all levels 2-6 digits). Approximately 57% of the table cells were sensitive according the to P% rule. For illustration and consistency with previous noise literature, the noise distribution used in all of our applications was the split triangular distribution discussed earlier relative to 10-20% changes. A P% value of 10 was also arbitrarily chosen for our analysis.

4.1 Random Noise

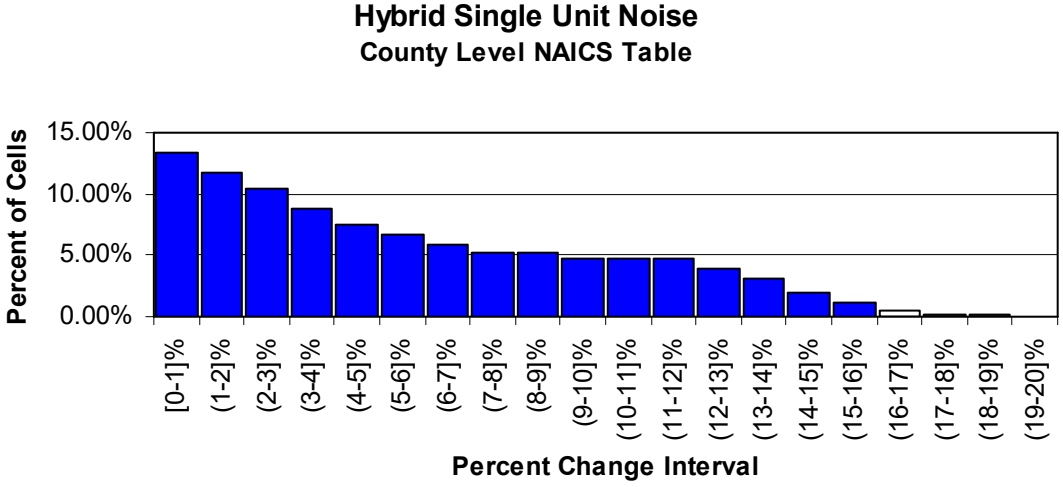
The graph below represents percent change intervals for the table of County by NAICS for a single state. Only cells with 3 or more respondents were included in this analysis since balancing was not performed on cells with less than 3 and those cells are by definition sensitive and we never want to intentionally balance them. As you can see, while there is some noise cancellation there is also a large number of cells that receive significant doses of noise. Keep in mind that there are many cells with 3 or more respondents that are still sensitive. These cells do tend to receive significant doses of noise and so help to form the chart's peak between 10% and 13%. Under this application of random noise, approximately 92.55% of all the sensitive cells received the full suggested protection from noise, according to the protection multiplier measure discussed earlier. The remaining sensitive cells each receive some portion of the noise suggested by the P% rule, and all sensitive cells are protected by the uncertainty about which cells changed in what direction and how much. Whether or not this level of protection is sufficient for a table is the responsibility of the appropriate authoritative body, such as the Census Bureau's Disclosure Review Board.



4.2 Completely Balanced Noise

To illustrate how balancing can improve results, the graph below represents the same table under completely balanced noise at the County by 6-Digit NAICS level. This subset of cells (assignment sub-table) represents the greatest amount of detail published for CBP, and can be considered the building blocks of all other published cells. The purpose of choosing this level of detail, as mentioned earlier, is in the hopes that the positive effects will 'trickle up' into the aggregate cells (consisting of all other table cells in this case). The issues and analysis supporting our assignment sub-table decision will be discussed in detail in a subsequent section.

In this example we treat each establishment as its own company, ignoring for now the Census requirement of company level protection for multi-unit firms. This technique gives us the most flexibility when balancing the noise. You can see a vast overall improvement compared to the graph representing random noise shown above, as there are many more cells receiving much less noise than before. We did not balance the assignment sub-table cells with less than 3 respondents as they are automatically sensitive, but we did balance any sensitive cell with 3 or more respondents. As such, the overall level of protection may have been compromised. The actual percent of fully protected sensitive cells was approximately 91.07%, showing a very minor decrease in protection compared to random noise. This minor reduction in protection pales in comparison to the great improvement in data quality.

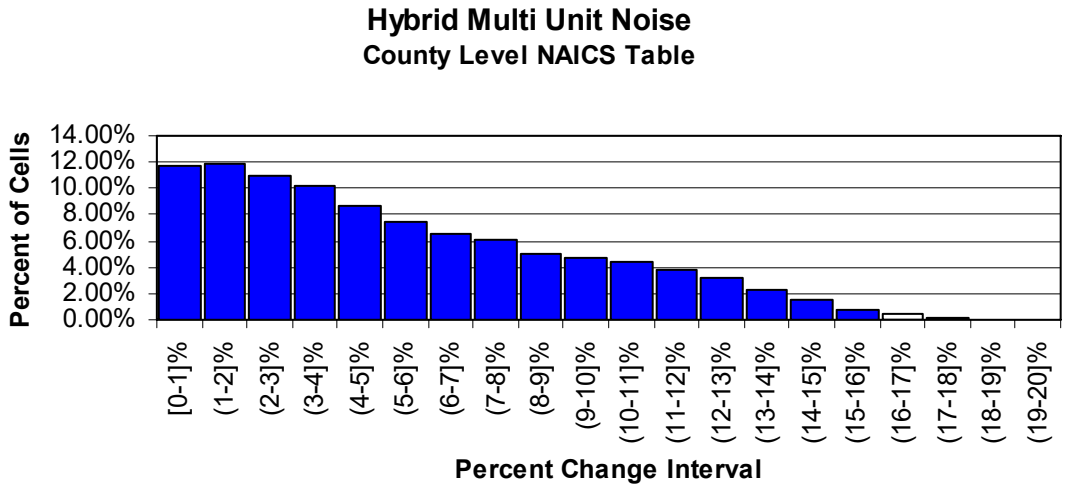


4.3 Balanced Noise with Multi-Unit Company Protection

To show that balancing is a practical option for business surveys that contain establishments from multi-unit companies, we also applied a modified version of the balancing algorithm to our data set that provides company level protection. This application uses a two step method for assigning noise factors;

- Assign factors to all establishments that belong to multi-unit companies randomly, ensuring that each establishment from any given company receives noise in the same direction.
- Assign factors to all single unit establishments applying the balancing algorithm to reduce noise, taking into account factors already assigned to multi-unit establishments.

Accounting for multi-unit company level protection reduces the amount of flexibility that balancing has to minimize noise since we cannot use multi-unit establishment values to balance noise. However, the graph below suggests that this reduction is extremely minimal, at least in the case of our CBP test data set. There is a slight reduction in the number of cells with small percent changes, but it still shows a radical improvement over random noise. It is reasonable to believe that the effect of company level protection on balancing is directly related to the proportion of establishment records that belong to multi-unit companies. In out test data set multi-unit companies accounted for more than 26% of the microdata records and more than 65% of the total payroll, which is likely typical of many business establishment censuses and surveys. The level of protection for this application of noise was 91.32%, a slight increase over single unit noise balancing that may be attributed to the slight reduction in the effects of balancing.



4.4 Choosing the Assignment Sub-Table

We had several options when determining what level of detail at which to balance the noise. We could have balanced at either the state or county level, and/or at any of the 6 hierarchal levels of NAICS industry codes. We decided that county level balancing would be more appropriate than state because there are many more county level cells than state level cells. We also expected a reduction in county level noise to carry up into state cells, which was supported by our testing and analysis. Which level of NAICS to use was slightly more complicated. The more detail we used the more cells would be affected due to the hierarchal nature of the coding, so we wanted to use the most detail possible. On the other hand, the aggregate cells are generally considered more important and so making sure that these cells are accurate took priority over the detailed cells.

In order to determine a balancing sub-table that provided a good compromise between our two objectives, we looked at the effects of balancing on the subsets of cells of different NAICS levels when noise was balanced on the various NAICS levels. The idea was to get a picture of how much the balancing effect 'trickled up' into aggregate cells or 'trickled down' into more detailed cells. Specifically we looked at the distribution of percentage change from noise of the various NAICS levels for the different noise applications. What this analysis showed was that there is a significant 'trickle up' and 'trickle down' effect for each level of balancing, and each is significantly better than random. More importantly, there did not seem to be a strong reduction in the effect of balancing on the sector level (NAICS 2 digit) estimates by balancing on more detailed cells. It was

this observation that led to our decision of balancing at the 6 digit NAICS level within county. This level of balancing affects the most cells possible and also ensures quality aggregate level estimates.

5 Rounding and Noise

Economic data is always published in some rounded form, often integers representing thousands or millions. This type of rounding can be done at the record level prior to any tabulation, or applied to the unrounded table values post-tabulation. Noise is designed to protect individual respondents by changing their response values by small percentages. Rounding can therefore systematically remove the effect of noise on small response values. In some cases this may not be an issue of concern, but under certain circumstances this could result in serious damage to the level of protection provided by noise. To deal with these situations we have investigated several rounding methods that could work to sustain the protection provided by noise. The methods we discuss involve rounding record level data prior to tabulation as well as others which are applied to table cell values.

Let us first discuss some general ideas related to rounding and introduce a sample data set that we will use to illustrate the methods we present. Noise is designed to protect continuous magnitude variables, although in reality all published numbers are rounded to some degree. As a result, we have to consider the case where small rounded numbers behave like discrete count values. It is impossible to change small integers by percentages to produce new integer values, and so we are forced to change them in integer amounts and accept the resulting percent change. It is worth noting that rounding itself provides some protection to response values. For example, consider applying noise to a survey in which the microdata values are already rounded to thousands. Assuming these values were reported in dollar amounts, small values have some uncertainty already built in. For example, any value between 500 and 1499 will be represented by a 1 (thousand).

To illustrate the effects of small value rounding on noise we look at a sample data set from the Census Bureau's Non-Employer Statistics Program. The data again represents a single state, and the table of interest consists of county level NAICS estimates (at various levels of industry) of total receipts values. More than 30% of all cells in this table are sensitive according to the arbitrary P% value of 10, and we apply balanced noise using a distribution of 10-20%. Here there are no multi-unit companies so we did not have to take that into account; however we have shown that this should not significantly affect the noise balancing process.

To establish a control for our investigation of rounding, we first examine our test table using unrounded noisy microdata values and unrounded cell values. This analysis represents what we would get if we could use and publish completely unrounded values. The table below represents the cell change distribution for all non-sensitive cells. We only look at non-sensitive cells because we expect (and want) sensitive cells to change significantly and are therefore more concerned with preserving the quality of the non-sensitive cells. From the table below the overall quality of these cells appears very high, as the non-sensitive cells are generally perturbed by very small amounts. Also, the level of protection for this table's sensitive cells is 89.23%. We consider this level sufficient and will use it as a basis for comparison; however any value such as this would need to be approved by the Census Bureau's Disclosure Review Board prior to publication.

Unrounded Values		
PERCENT CHANGE	**COUNT**	**PERCENT**
0-1%	18941	73.99%
1-2%	1898	7.41%
2-3%	1260	4.92%
3-4%	904	3.53%
4-5%	706	2.76%
5-10%	1601	6.25%
10-15%	288	1.13%
15-20%	2	0.01%
20% +	0	0.00%

5.1 Record Level Rounding

There are several reasons that a survey program may want to work only with rounded noisy microdata values. Since many programs will be applying noise to already rounded sample values, it may not make sense to produce noisy values with a higher level of detail. Also, in order to publish tables that are rounded, they may require that the microdata be rounded to the same degree in order to maintain additivity.

Let's say for example that the minimum noise factor was 10% for any given record. That means that all values less than 5 will never change as the result of noise. This may not be a problem for certain tables or certain types of surveys, but if the data contain many small microdata values and the published tables contain many cells that are based on only a few contributors it could be a serious problem. That would guarantee that any cell that is less than 5, and many that are significantly larger, will represent the true response values unaffected by noise. To address this problem we may want to ensure that all response values change by at least some amount after rounding, or possibly that all values had some positive probability of changing.

5.1.1 Standard Rounding of Microdata Values. In order to demonstrate the negative effects of simply rounding noisy values at the record level, we applied this procedure to the sample Non-Employer data set. The resulting level of protection for our test table was reduced to 79.09%. Initially this may not seem like a significant decrease in protection, but with a closer look you will find that the subset of sensitive cells with values less than 10 goes from a protection level of 96.30% to 56.09%. This shows that the protection provided by noise to these cells is systematically reduced by the standard rounding procedure. It is therefore in the best interest of the program to consider alternative rounding schemes that may help to retain this protection.

5.1.2 Ceiling/Floor Rounding for Small Response Values. The ceiling/floor rounding scheme is very simple;

- If a noisy decimal value is greater than the original cell value, then round the noisy value up to the next larger integer (or whatever rounding level desired).
- If a noisy decimal value is less than the original cell value, then round the noisy value down to the next smaller integer.

This method can be applied to all response values, just the ones that do not change naturally as a result of noise, or some other subset of small values. The noise factor distribution controls how much a value changes by controlling the parameters that describe the distribution. The ceiling/floor rounding function ensures that every value will change by at least 1 unit after rounding, regardless what percentage change that represents. We applied the ceiling/floor rounding function to every value in our data set, and achieved an overall protection level of 89.23% and 96.30% for sensitive cells less than 10. This is approximately equal to that of the unrounded data. Since this method always introduces more change than pure EZS Noise, it is necessary to ensure that the data quality has not been compromised. In our test table there was no apparent decrease in data quality compared to the unrounded values, as shown in the table below, and therefore this method becomes a very attractive option for microdata level rounding. The only difference when compared to the unrounded values is that there are a few cells that receive larger amounts of noise (20% +); however these changes can be attributed mostly to small values that change by small integer amounts relative to large % changes.

Ceiling/Floor Microdata Rounding

PERCENT CHANGE	COUNT	PERCENT
0-1%	19057	74.44%
1-2%	1919	7.50%
2-3%	1300	5.08%
3-4%	901	3.52%
4-5%	683	2.67%
5-10%	1317	5.14%
10-15%	303	1.18%
15-20%	74	0.29%
20% +	46	0.18%

5.1.3 Probability Rounding for Small Values. Another option for rounding small response values is to round such that only some of the small values change, and that others are allowed to remain the same. In the same sense that noise provides protection to response values by creating uncertainty about how much or which direction a value was changed, in this way it can also create uncertainty about whether a value has changed or not. This modification of the noise method starts to become similar to synthetic data methods designed to handle categorical or more specifically ordinal response variables. Using our standard noise distribution (10-20%), values less than 5 do not always change, and 1's and 2's never change. In our application of probability rounding we applied random change probabilities of .40, .55, .70, and .85 for values of 1, 2, 3 and 4 respectively. The protection levels obtained under this application were 80.33% overall and 55.56% for sensitive cells less than 10. This method does not appear to produce a significant improvement over standard rounding; however may still have some benefit. In the case of a survey for which there is only moderate concern about small sensitive cell values, it may be sufficient to use a method for which each such value **could** have changed. As such, users cannot assume that all small values are accurate. This could be particularly useful when a survey does not produce tables with many small sensitive cell values.

5.2 Table Level Rounding

Another rounding option is to use the decimal noisy values to create the tables to be published, and then apply a rounding scheme to the cell values at the table level. The drawback to this method is that additivity of the table usually does not hold if aggregate cells are derived from microdata records rather than the rounded interior cell values. On the other hand the benefit is that you have more control over how much cells are distorted when working at the table level. Standard rounding of these cell values may not cause any severe protection problems; however we have found that in the case where there are many small response values it does. Sensitive cells that are derived solely or primarily from small values are not likely to receive enough noise. Applying standard table level rounding to our test table produced an overall protection level of 80.54%, and 55.56% for small sensitive values, results which are very similar to those produced by standard rounding at the microdata level.

5.2.1 Ceiling/Floor Rounding for Cell Values. In the same manner that we treated microdata response values earlier, you can simply ensure that each cell value changes by at least 1 by rounding up or down according to the direction of the overall cell noise. Again, this can be applied to every cell or just a subset such as all small cells. This method can work very well to both maintain protection to small sensitive cells and minimize the reduction of data quality that a rounding technique can add. We applied this method to every cell in our table and achieved an overall protection level of 92.75%, and 100% for sensitive cells with values less than 10. This is obviously the best method as far as level of protection provided, and produces similar data quality results to those obtained with microdata level rounding. Again there are a handful of cells that change by large percentages, as shown in the table below, but these changes can be attributed to small cell values forced to change by small integer amounts resulting in large percentage changes.

Ceiling/Floor Cell Rounding

PERCENT CHANGE	COUNT	PERCENT
0-1%	15790	61.68%
1-2%	2766	10.80%
2-3%	1763	6.89%
3-4%	1223	4.78%
4-5%	910	3.55%
5-10%	2338	9.13%
10-15%	649	2.54%
15-20%	104	0.41%
20% +	57	0.22%

6 Conclusions

In this paper, we have extended the basic EZS Noise method to produce a version that we call Balanced EZS Noise. We then showed how noise interacts with various rounding methods. We described the several advantages that the basic version of the EZS method has over cell suppression and other related protection methods for protecting magnitude data tables. Although this basic version (now referred to as Random EZS) is suitable for many surveys, we have discovered during our work with unweighted data that the EZS method can easily be improved. Balanced EZS noise can greatly improve data

quality in a selected key table, does not harm data quality in other tables, and overall does little harm to the degree of protection (from disclosure) in all tables. Balanced noise, although somewhat more complicated to implement than Random EZS, is still much simpler than cell suppression and related methods while retaining all the key advantages of Random EZS. The interaction of rounding and noise is quite important. Rounding has different goals than noise, but they both perturb the data and therefore both impact both the data quality and the level of uncertainty about the underlying microdata. The problem is that in certain situations the perturbations from rounding can "undo" the perturbations from noise, thereby eliminating the protection provided by noise. We analyzed some rounding methods and showed how they can be modified to work well with noise under different circumstances.

References

1. (CFS05) Commodity Flow Survey (CFS) Conference (2005), July 8-9, Boston, Massachusetts. Participant research questions at: http://www.trb.org/conferences/cfs/Workshop-DataProducts-Question.pdf

2. Dula, Jose H.; James T. Fagan, Paul B. Massell, (2004) "Tabular Statistical Disclosure Control:Optimization Techniques in Suppression and Controlled Tabular Adjustment" http://www.census.gov/srd/papers/pdf/rrs2004-04.pdf

3. Evans, B. Timothy (1997), "Effects on Trend Statistics of the Use of Multiplicative Noise For Disclosure Limitation", ASA Proceedings of the Section on Government Statistics.

4. Evans, Timothy, Laura Zayatz, John Slanta (1998), "Using Noise for Disclosure Limitation of Establishment Tabular Data", Journal of Official Statistics http://www.jos.nu/Articles/abstract.asp?article=144537

5. Massell, Paul B. (2005), "The Interaction of Noise and Weighting in Protecting Company Data from Disclosure", unpublished note.

6. Massell, Paul B. (2006), "Using Uncertainty Intervals to Analyze Confidentiality Rules for Magnitude Data in Tables", http://www.census.gov/srd/papers/pdf/rrs2006-04.pdf

7. Massell, Paul; Zayatz, Laura; Funk, Jeremy; (2006) Protecting the Confidentiality of Survey Tabular Data by Adding Noise to the Underlying Microdata: Application to the Commodity Flow Survey, appears in: Josep Domingo-Ferrer, Luisa Franconi (Eds.) :Privacy in Statistical Databases, CENEX-SDS Project International Conference, PSD 2006, Proceedings. Lecture Notes in Computer Science (LNCS) 4302, Springer 2006, ISBN 3-540-49330-1.

8. Wolter, Kirk (1985), Introduction to Variance Estimation, Springer,

9. (WP22) Federal Committee on Statistical Methodology (FCSM) (revised 2005), Working Paper 22, http://www.fcsm.gov/working-papers/spwp22.html

10. Zayatz, Laura (2000), "How rounding should be incorporated into the p% rule", unpublished note.

11. Island Areas Census program at U.S. Census Bureau http://www.census.gov/population/www/proas/pr_ia_ecen.html

12. Non-Employer Statistcs Program at U.S. Census Bureau
http://www.census.gov/prod/2006pubs/ns0400a01.pdf

13. Wu, Jeremy S.; Abowd, John M., "LEHD and Noise Infusion", http://lehd.dsd.census.gov/led/library/presentations/PN-2006-05.pdf

Appendix – Noise Factor Assignment

Random Noise Factor Assignment

- Each Company is randomly assigned a direction, (+/-)

- Each Establishment level record is then assigned a random noise factor based upon its respective company direction
    - If the direction is +, then the factor is randomly generated from the right half of the split triangular distribution; i.e., the portion greater than 1
    - If the direction is -, then the factor is randomly generated from the left half of the split triangular distribution; i.e., the portion less than 1
    - Note that Single Unit Companies each have one establishment level record, while Multi Unit Companies can have many

- Factor Assignment for New Records in Subsequent Years
    - If a new establishment belongs to a multi-unit company that is already present in the data set, it is assigned a random noise factor using the previously assigned noise direction for its company
    - If the establishment is a single-unit or new multi-unit, it is assigned a new factor according to the procedure above

Balanced (Hybrid) Noise Factor Assignment

Stage 1 - Multi-Unit Companies

- The subset of all records belonging to multi-unit companies are assigned factors first
    - To do this they are physically separated from the complete file

- These records are assigned random noise factors according to the procedure described above

Stage 2 - Single-Unit Companies

- The multi-unit and single-unit records are again combined, with multi-unit establishment factors intact

- These records are sorted according to the assignment table definition
    - The assignment table is a single table for which it is considered important to maintain data quality
    - It may also be at a low level of hierarchy so that any balancing effects will also improve aggregate estimates

- Within these assignment table cells, records are sorted according to
    - Multi-Unit records first, then Single-Unit records
    - Descending size within Single-Unit records

- Factors are assigned to records in one assignment table cell at a time, and to each record within a cell according to the above designated order
    - For each record, the total cell noise distortion is calculated by adding the net distortion from the current record to the sum of the distortions from all records within that cell with previously computed noise factors
    - In this way the total cell distortion resulting from any multi-unit records in a cell (which already have factors assigned) can be determined prior to assigning single unit factors

- Starting with the largest single-unit respondent in a cell, the noise direction is assigned based on the total cell distortion thus far (running noise total)
    - If there are no multi-unit records in an assignment cell (zero net noise distortion) then this direction is assigned randomly
    - If the net distortion is +, corresponding to a + noise direction, then this respondent will get a – noise direction, and visa versa

    - This respondent's noise distortion is then added to the net cell distortion
    - This process is then repeated for all remaining single-unit respondents in the cell

- If there are only 2 companies in a cell, and at least one is a single unit record, then it is treated slightly differently (this is the 'hybrid' portion of the method)
    - Single-Unit records in 2 company cells are all assigned random noise directions
    - These cells are by definition sensitive, so there is no reason to balance
    - This constitutes the 'hybrid' portion of the method, and significantly increases the level of protection provided by noise

Assigning New Factors in Subsequent Years

- If a new record belongs to an existing multi-unit company (has a noise direction assigned in a previous year) then it is assigned a random noise factor generated using its company's respective noise direction

- All new single-unit establishment records and multi-unit company records are assigned noise factors according to a balancing procedure similar to what was used in the initial assignment of noise
    - Multi-unit records are assigned random noise as discussed above
    - Records are then combined and sorted according to desired balancing table
        - ALL records with factors already assigned are listed first in a cell, with any new records following after
        - This is similar to the initial balance sorting, where multi-unit records are the only ones with factors already assigned
    - Net cell distortion is summed from the records with factors already assigned, and new records are assigned factors based upon this net distortion