Protecting the Confidentiality of Tables by Adding Noise to the Underlying Microdata

Paul B. Massell, Jeremy M. Funk, both at the U.S. Census Bureau

Abstract

Users of statistical tables released by the Economic Directorate of the U.S. Census Bureau have raised the issue of whether an alternative to cell suppression can be used to protect the confidentiality of such tables. These users would like to have access to at least an approximate value for each cell, except possibly for those cells that are the most sensitive. An alternative method was developed several years ago by researchers at the Census Bureau that successfully meets that goal. This method uses a carefully calibrated noise distribution to generate noise which is then added to the microdata values of a magnitude variable requiring protection. These noisy microdata values are then tabulated to form the cell values for all the tables in a statistical program that describe that variable (e.g., receipts for Non-Employer Statistics). This method is conceptually simple and easy to implement; in particular, it is much simpler than cell suppression. The main concerns are whether noise protected tables are fully protected and whether the noisy cell values are as or more useful to users than the combination of exact and suppressed values provided by cell suppression. The seminal paper by Evans-Zayatz-Slanta (J. Official Statistics, 1998) showed that this was clearly true for the survey analyzed in that paper. The work presented in this paper provides analysis for additional surveys with different features than the survey described in the earlier paper. We present general protection arguments that involve ways of relating the uncertainty provided by noisy values to the required amount of protection. We present graphs which show the different distributions of net noise on the set of sensitive cells versus that for the non-sensitive cells. We also discuss some ways to fine-tune the algorithm to a particular table, taking advantage of its special characteristics. Our conclusion is that this method works well for tables from a wide variety of statistical programs within the Economic Directorate.

Keywords: Disclosure Avoidance, Confidentiality, Protection Methods, Noise

1. General Comparison of EZS Noise with Deterministic Protection Methods

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 who is able to write a linear programming program, could write an audit program which computes 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 respondent’s contribution. We claim that such wide intervals are common in practice. More realistically, most users do not 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.

1 This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed on statistical issues are those of the authors and not necessarily those of the U.S. Census Bureau.

2 The numerical results in this paper are not intended to correspond to officially released data. The type of noise protection and the specific noise distributions used to generate noise for the microdata underlying the tables that were analyzed in this paper may or may not correspond to what will be used in production runs.
We use the expression ‘deterministic protection methods’ to refer to a set of protection methods that are 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.

For the EZS noise method the determination of required uncertainty is still done at the cell level, but in contrast to cell suppression, uncertainty is not created there. In EZS noise, the values of cells (both sensitive and safe) are perturbed indirectly via the underlying microdata contributions to the 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]).

Other Problems with Cell Suppression

We have mentioned the basic problem with cell suppression from the table user's point of view; the suppressed cells don't reveal any information. In cases where a cell suppression pattern contains a high ratio of secondary to primary suppressions, both users and the office itself may feel that too much non-sensitive cell value information (secondary cell suppressions) is being wasted simply to protect the sensitive cells.

From the office's point of view there are additional problems. There is the issue of software development. If an office wants to develop its own computer program for cell suppression, it is likely to require the skills of an advanced scientific programmer. The network flow algorithm is extremely fast, however this computational algorithm is guaranteed to work only for two dimensional tables. By “work” we mean the algorithm computes a suppression pattern which fully protects every sensitive cell. For three or higher dimensional tables, network flow cannot guarantee a fully protective pattern. In order to guarantee a fully protective suppression pattern one must use a search code that is based instead on a general linear programming (LP) model. In addition, even when this LP code is well-written it can be very time-consuming to run on large tables. If the office would like to minimize runtimes the cost of the required software can be significant. The EZS noise method is typically much faster for any given class of tables and requires no specialized software.

Another 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 ‘backtracking’. 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 the ‘unraveling process.’

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 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).
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 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. 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.

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 will be a valuable measure of the effect noise has on data quality.

After mentioning several ways in which EZS noise is better than cell suppression, a natural question is whether cell suppression is better than EZS noise in any way? The answer is perhaps, and has to do with the perception of users to a table protected with EZS noise. For example suppose an office decides to release all cells in a table, including the sensitive ones. The noise distribution used has a left component that is nonzero on [0.80, 0.90] and a right component that is nonzero on [1.10, 1.20]. Take a cell with two contributors with values of 50 and 100. If the 50 value has a noise factor of 0.80 and the 100 has a noise factor of 1.10 then the resulting noisy cell value is 150, the same as the sum of the pre-noisy values. If one of the contributors ignores the information about the noise procedure that was applied to the data, he will subtract the value he provided to the office from the released total and derive the exact value of the other contributor. Thus in an EZS protected table, if the user ignores the noise information, he may sometimes be able to derive values that are very close to the true values. The office has to decide whether to regard this problem as serious or not. If the office does not regard recovery of values based on faulty assumptions as disclosures, then the office would consider the contributions to be fully protected.

2. 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 with 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 much of our work, and in our examples below, we use either the standard or extended p% rule for determining both the sensitivity status of each cell and the desirable amount of perturbation. Simply put, the p% rule determines an amount of suggested perturbation 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 errors, 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 be always be small enough to meet the SO’s error limits or will meet these limits a high percentage of the time. 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 the ‘calibration’ of the noise distribution.

There are a variety of reasonable ways that the two basic types of error can be measured. Below we define an “alpha error” that measures under-perturbation of sensitive cells, and a “beta error” that measures over-perturbation of both safe and sensitive cells. These are “global scalar measures” in the sense that for each type of error, there is a single value for a given table. In order for a SO to use these measures for deciding when a level of noise is too high or too low, it would need computational experiments with a variety of noise distributions to determine acceptably small values for alpha and beta.

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

Alpha Error: A Global Scalar Measure of Under-Perturbation

\[
\text{Alpha Error} = \frac{\sum_{\text{All Sensitive Cells}} \max \left\{ 0, 1 - \left( \frac{\text{Actual Perturbation}}{\text{Nominal Perturbation}} \right) \right\}}{\text{Number of Sensitive Cells}}
\]

Beta Error: A Global Scalar Measure of Over-Perturbation

\[
\text{Beta Error} = \frac{\sum_{\text{All Safe Cells}} \left| \frac{X - Y}{X} \right| + \sum_{\text{All Protected Sensitive Cells}} \left| \frac{X - Y - \text{prot}}{X} \right|}{\text{# Safe Cells} + \text{# Protected Sensitive Cells}}
\]
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 the protection multiplier (PM), defined as the ratio of the absolute value of the perturbation to the protection suggested by the p% rule,

\[
PM = \frac{\text{Perturbation from Noise}}{\text{Suggested Perturbation}}
\]

It is useful to compute the percentage of all sensitive cells that have PM \( \geq 1 \). 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 < 1. Of course, such suppressed cells may be recoverable, unless the SO takes the computationally costly step of running a complementary suppression program and applies it to those cells.

**The Percentage Change Distribution: A Global Measure of Noise’s Effect on Data Quality**

The distribution of percent changes to a designated subset of safe 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 a small percentage, say between 0 and 3%. If a significant percentage have been changed by more than say 5%, the SO can try to lower the amount of noise being added by changing the noise distribution or may consider using the balancing version of the EZS method described below.

**EZS Noise Factors for Unweighted Data**

- Let \( X \) = original microdata value
- Let \( Y \) = perturbed value
- Let \( M \) = noise multiplier (or factor); i.e. a draw from a specified noise distribution

\[
Y = X \times M
\]

The distribution we used for our examples below is the “split” triangular distribution, for which the density function is described below and illustrated in figure 1.

\[
\begin{align*}
\text{for } & \quad 2-b < x < 2-a, \quad f(x) = k \cdot (x - (2-b)) \\
\text{for } & \quad a < x < b, \quad f(x) = (-k) \cdot (x - b) \\
\text{otherwise } & \quad f(x) = 0
\end{align*}
\]

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

**3. Using EZS Noise to protect Tables generated from Weighted Microdata**

**Example: Commodity Flow Survey**

Microdata collected as part of a sample survey is typically weighted in order to produce an estimate of actual population values, so as an example of how noise works with this type of data we look at the application to the Commodity Flow Survey (CFS). This survey is more complex than many economic surveys because there are two stages of sampling and a total of seven

**Split Triangular Density Function**

![Split Triangular Density Function](image)

**Figure 1: The Split Triangular Distribution**

The density is piecewise linear and is symmetric about 1. In our examples, we use \( a = 1.05 \) and \( b = 1.15 \).
weights and adjustment factors. After giving a brief introduction to the CFS, we will discuss several noise features which would apply to any weighted survey data and the results of our application of noise to the CFS.

Introduction to CFS

CFS data provide information on shipments originating from manufacturing, mining, wholesale, auxiliary warehouses, and selected retail establishments in the 50 states and the District of Columbia. While it is in fact an establishment survey, its focus is on the characteristics of shipments rather than the establishments or companies themselves. The survey’s goals are to estimate the characteristics associated with the origins and ultimate destinations of shipments, the distances traveled by these goods, the types of commodities shipped, the modes of transportation, and the volume of shipments measured by weight and value.

Issues Related to Weighted Data

The purpose of the EZS noise method is the protection of tabular data; however there are various other sources of uncertainty that also provide protection as a positive side effect. The use of noise is essentially the application of an additional weight that distorts the actual reported record level value. Sampling weights can have the same effect on data, and so this additional source of uncertainty should be accounted for when determining how sensitive a cell is and how much noise a record needs.

In some of our test tables, sampling weights fully protect many of the one and two-company cells. This is in contrast to census data (or more precisely, data for which the weights always equal 1) for which any cell that consists of data from only one or two companies is always sensitive and thus requires protection. As a result, survey data in which most records have weights larger than \(1 + \frac{p}{100}\) are likely to have few sensitive cells when sensitivity is determined using the extended p\% rule (a version of the p\% rule that takes into account protection from sampling weights).

Distinguishing Known Weights from Unknown Weights

CFS tables are protected by seven weights and scaling factors that are applied to the microdata. Four of these weights are generally unknown to data users, but three of these weights are often public knowledge. Because weights that are known by potential data intruders could be used to improve an estimate of a respondent’s value, these weights are not considered to provide protection. One of the simplest scaling factors is the one that multiplies an estimate of a week’s worth of shipments to estimate a quarter’s worth. This multiplier is 13 and is the same for all microdata. It is easy to see that this type of multiplier could be known to many table users and since there is no uncertainty associated with it, it would provide no disclosure protection.

As an example of an unknown weight, consider the sampling weight (for a given microdata value). To be more precise about its being ‘unknown’, when a sampling weight equals 1, some users may know that the company had that weight, i.e., the company was selected for the sample with “certainty”. However, when a sampling weight is greater than 1, the uncertainty that even well-informed table users have about its value, is a function of its distance above 1. We make this precise below when we introduce the notion of an uncertainty model for an unknown weight.

The distinction between known and unknown weights is needed when applying the extended p\% rule to determine protection needs for a cell. In that rule, as described in (WP22, [9]) one computes the remainder (rem) as follows:

\[
\text{rem} = T - (X_1 + X_2)
\]

In which \(T\) is the fully weighted estimate (i.e. all weights are used to compute \(T\)) but \(X_1\) and \(X_2\) are unweighted microdata values. However, if there are known weights, they should be applied to \(X_1\) and \(X_2\). That is

\[
\text{rem} = T - (X_1 \cdot w_{kn1} + X_2 \cdot w_{kn2})
\]

where \(w_{kn1}\) and \(w_{kn2}\) represent the product of all known weights for the given microdata values. The formula for the suggested perturbation ‘prot’ also needs to be generalized:

\[
\text{prot} = ((p/100) \cdot X_1 \cdot w_{kn1}) - \text{rem}
\]

When all weights are unknown there is a simple way to construct a multiplier that combines the noise factor and these weights (Evan, Zayatz, Slanta, [3]). Let ‘\(w_{un}\)’ be the product of all the unknown weights

\[
Y = X \cdot (M + (w_{un} - 1))
\]
The use of this formula when applying noise means that noise factors will have less impact on values that are heavily weighted. This is desirable since the uncertainty about unknown weights is usually significant. It is necessary to assume an uncertainty model that makes our assumptions precise. We assume that even the best informed table users know only that an unknown weight ‘w’ lies somewhere in the interval \([w - k*(w-1), w + k*(w-1)]\) where \(k \geq b\), and where ‘1+b’ is the upper noise limit. One can show that if such an uncertainty model holds for each individual unknown weight, it holds approximately for their product \(w_{un}\). With just a few lines of algebra, one can also show that this uncertainty model for \(w_{un}\) implies that the full noise-weight multiplier has an uncertainty of at least \([1-b, 1+b]\), which is the uncertainty range for the noise factor. That is, if the unknown weights satisfy our uncertainty model, the noise-weight multiplier provides as much uncertainty for weighted data as the noise factor provides for unweighted data.

When there are both known weights and unknown weights, this formula needs to be generalized a bit further. Let \(w_{kn}\) be the product of all the known weights. Then,

\[
Y = X * (w_{kn} * (M + (w_{un} - 1)))
\]

Results of CFS Application

The graphs below represent the amount of change that cells in the CFS Origin State by Destination State by 2-Digit Commodity table sustained as a result of the noise addition. This table consisted of 61,174 cells, only 230 of which are sensitive. Note that the majority of sensitive cells receive noise in the range 5% to 11%, while the non-sensitive cells tend to receive very little or no noise at all.

From the above graphs you can see that overall nonsensitive cells are not receiving large amounts of noise, and therefore the utility of the published data is being preserved quite well. Although we see that sensitive cells do tend to receive significant doses of noise, it is necessary to more precisely measure the amount of protection provided by the noise. To do this we calculated a protection multiplier (PM) for every sensitive cell. PMs are defined as the ratio of actual perturbation from noise divided by the amount of perturbation required by the p% rule. If this ratio is greater than 1 for a sensitive cell, it is considered to have been fully protected by noise. In the CFS table discussed here, only 2 of 230 sensitive cells fail to receive the perturbation required and thus have PMs less than 1.

Calibration: Finding a Good Noise Distribution to Use for EZS

Finding a good noise distribution for a particular application of the EZS method may require some calibration. The way that noise will affect a particular survey is a function of the characteristics of the microdata, the characteristics of the set of tables to be generated from it, and the manner in which noise is applied to the data. Because this result is highly data dependent, it is not always clear ahead of time what the best noise distribution may be for a particular survey. In the example presented here, we found the amount of distortion to non-sensitive cells to be acceptably low and the amount of protection to sensitive cells to be sufficient. One could, however, experiment with the magnitude of the noise distribution to find the minimal amount of noise that provides sufficient protection, thus minimizing the amount of damage done to tabular data quality.
4. Using EZS Noise to Protect Tables Generated from Unweighted Census Data, Example: Non-Employer Statistics

Applying noise to unweighted data introduces some new issues. The lack of weights means that the pure noise factors are applied directly to the respondent values and there is a direct relationship between the magnitude of the noise distribution and amount of cell value distortion. This also means that the amount of protection required for a sensitive cell is not influenced by weights. Therefore the amount of protection required for single respondent cells under the p% sensitivity rule is always exactly p% of the total cell value. Similarly with two respondent cells, the amount of protection required is always exactly p% of the larger respondent’s value.

This relationship directly influences the choice of noise distribution used to generate factors. It is a simple measure to set the minimum value of the noise distribution equal to 1 + p, which ensures that all single respondent cells will automatically receive full protection. Whether or not multiple respondent cells receive full protection is a function of both the randomness of the noise factors and the distribution of respondent contribution sizes within a cell.

To examine the effect of EZS noise on unweighted microdata we will now look at some tables produced by the Non-Employer Statistics (NE) Program at the U.S. Census Bureau. The example table discussed here represents county by detailed industry code for one U.S. state. It consisted of 72,527 cells, 24,607 of which are sensitive. The same distribution that was used earlier for CFS data was used here for NE, reflecting a minimum change of 5% and a maximum change of 15%. The graphs below show the distributions of percent change to non-sensitive and sensitive cells.

Comparing the graph of non-sensitive cells to the one presented earlier for CFS, it is immediately clear that the lack of weights has caused the noise to have a significantly stronger impact on the non-sensitive cell values. The natural effect of the noise to cancel itself out is still apparent as the majority of these cells are still receiving smaller amounts of noise; however this effect is less pronounced.

In order to assess how well the sensitive cells are being protected, we again look at the distribution of PMs. There are many more sensitive cells in this table when compared to the CFS example, but this is easily explained by the lack of weights and detailed nature of the table; there are many one and two respondent cells in NE tables. The shape of the graph above looks much like the actual noise distribution, which again can be explained by the prevalence of single respondent cells. Probably the most important statistic to look at is the percent of sensitive cells that receive full protection form noise (PM >= 1), which is more than 92% for this table.

5. Balancing Noise at the Cell Level to Preserve Data Quality

A natural question to ask when considering using noise as a disclosure avoidance technique is whether or not there is a way to add less noise to the non-sensitive cells, while not sacrificing the amount of protection provided to the sensitive cells. The primary benefit of EZS noise over cell suppression is that noise may allow for the release of more usable information in the form of noisy cell values. Suppression has the advantage that all published cell values are the actual estimates derived from the survey, so if the noisy cell values are perturbed too much then the benefit of noise over suppression is significantly reduced. As a result we have investigated various methods that may be able to reduce the overall amount of noise added to the data without compromising the level of protection.
One such method is to balance the noise at the cell level in a single selected table. The method assigns noise factors to the records in a cell in such a way that it works to minimize the final total noise in that cell. To achieve this we use a “greedy algorithm” that seeks to minimize the running noise total at each stage of balancing by selecting the direction of the next noise factor to be the opposite of the sign of the running total within that cell. This does not completely eliminate the noise, since factors are still randomly generated from the selected side of the noise distribution, but it does significantly reduce noise to cells with several contributors.

The main problem that arises with the use of this method is determining the set of cells to use when assigning the noise factors. One of the premises of the EZS method is that each record in a microdata set receives one noise factor, and all tables produced are protected using that set of factors. Ease of implementation is also an important and attractive feature of EZS, and so we would like to retain as much simplicity as possible. To keep the method simple we need to be able to define a set of cells such that each respondent corresponds to exactly one cell in the set (we call this set an ‘assignment sub-table’). In this way we can assign noise factors to every respondent just once as we traverse this set of cells.

Ideally we would like to be able to assign a single factor to each record that will provide the maximum benefit to all tables produced from the data. An algorithm may exist that would be able to determine the optimum pattern of noise direction allocation for situations like this, but such a method would certainly increase the complexity of the EZS method significantly. The solution that we have developed is a compromise because it allows for noise to be ‘balanced’ only for a single table. Therefore, we have to consider the effects on all other tables (generated from the same microdata) including aggregates and marginal values, and separate tables with different structures. In hierarchical tables, we expect that balancing the noise in cells of a high level of detail (e.g., the interior cells) will ‘trickle up’ into the aggregate cells. This is based on the basic idea that summing cells with less noise should in turn produce aggregate cells with less noise. Also, if certain tables are different but highly correlated, balancing the noise in one may have a positive effect on the other.

In the example of noise addition to unweighted NE data, the absence of weights resulted in a significant amount of distortion to the non-sensitive cells, making it a good example to look at the effectiveness of such a balancing scheme. In NE microdata, each establishment record can be viewed also as a company record because (almost) all NE companies are single-unit. This data feature leads to the simplest form of the balancing algorithm. We have only to choose a single table and a set of cells within it on which to balance (called a ‘balancing’ sub-table).

The structure of the NE tables also plays a significant role in determining how well the balancing of noise may work in this application. Published tables are at the national, state, and county level by industry classification (NAICS), which is completely hierarchical. This means that it would be possible to look only at the cells with the greatest level of detail; these interior cells can be considered the ‘building blocks’ of all other cells in the hierarchical structure. In this case these cells are the county by detailed industry code. A nice feature of the NE table structure is that even though different levels of industry detail are published for different industries and sub-sectors, the levels are clearly defined so that a list of the ‘most detailed published industry codes’ can be defined. Each respondent is assigned exactly one of these codes.

In order to determine whether a noise balancing scheme is beneficial to NE data tables, we looked at two main issues:

1. Does balancing have a strong positive effect on reducing distortion to non-sensitive table estimates?
2. Does balancing significantly reduce the level of protection provided by noise?

EZS noise was applied and balanced at several levels of detail, including various levels of industry and geography. One special condition we applied was that if a cell only had two respondents, noise factors were generated randomly instead of being balanced. Since these cells are automatically sensitive, reducing their noise level via balancing is not appropriate. We found there was a significant reduction in the amount of noise to non-sensitive cells for all tables on which we tested balancing. There was also a strong ‘trickle up’ effect of balancing; i.e., low detail cells (e.g., aggregates) benefited from balancing interior cells (or other cells with a high level of detail). We decided the best set of cells to which to apply noise were the interior cells; i.e. cells with the highest level of detail. To illustrate our
results, the tables below compare the distributions of noise to table cells for this particular balanced table for the same NE test state discussed in section 4.

![Graph of non-sensitive cells]

The graph of non-sensitive cells above shows a significant decrease in the amount of noise applied when compared to the earlier example of random noise application. The distribution of noise to sensitive cells remains very similar, with only what appears to be a slight increase in the number of cells that receive smaller amounts of noise. This significant decrease in excess distortion to non-sensitive cells seems to outweigh the reduction in noise, and possibly amount of protection, to sensitive cells. In this NE test table approximately 7.9% of sensitive cells receive less than full protection under random noise, while approximately 8.3% receive less than full protection under balanced noise. There is a relatively small increase in the number of cells that do not receive full protection from noise, and the increase in data quality of the resulting tables certainly warrants the use of balancing here.

The success of noise balancing to individual NE tables, the prevalence of the ‘trickle up’ effect in the hierarchical structure, and the minimal reduction in protection, makes the NE tables an excellent candidate for the incorporation of this balancing method into the EZS noise procedure. While these results are applicable only to this data, it is likely that they will also apply to other similar data products. Surveys that have certain special characteristics such as single-unit response data and a key table to which others are simply related, may greatly benefit from the application of this type of noise balancing.

## 6. Conclusions

EZS noise has several advantages over cell suppression for protecting economic magnitude data tables. Some of these advantages, e.g., the ability to protect linked tables easily and consistently, are also advantages with respect to other perturbative methods that are based on the same underlying algorithms as cell suppression. In addition, EZS noise generally has the advantage of simplicity of implementation compared to cell suppression and related perturbative methods.

After reviewing the basic features of the original version of the EZS, we showed how EZS could be generalized a bit to handle a survey such as the Commodity Flow Survey with several weights some of which are known to users and some unknown. This set of weights creates tables with few sensitive cells. These few sensitive cells tend to have respondent values with small unknown weights (i.e. close to 1). For most microdata values, the noise contribution to the joint noise weight multiplier is small compared to the contribution from unknown (to users) weights. Thus the percentage change of the cell estimates due to noise is often very small in cells dominated by such respondent values. In other words, when unknown weights add a lot of uncertainty to respondent values, little noise is needed to protect these values.

We also discussed the application of the basic EZS method to unweighted microdata from the Non-Employers Statistics program. In this program there is one key table, for which a large percentage of its cells are sensitive. (Recall for unweighted data, cells with only 1 or 2 contributors are sensitive.) It is necessary to add significant noise to the sensitive cells, but when that is done in the standard EZS method, this level of noise creates an unacceptable percentage change distribution for the safe cells. The solution is to reduce the level of noise in cells with 3 or more contributors since these cells are usually non-sensitive. The reduction of noise is achieved via a simple balancing algorithm. This method works to minimize the amount of noise in each cell in a selected set of cells that we call the ‘balancing sub-table’. This minimization is achieved by selecting noise directions that maximize noise cancellation. This sub-table has the property that
each microdata record appears in at most one cell of this sub-table. When the balancing sub-table is selected carefully, the effect is to greatly reduce the median level of noise to all cells in the balancing sub-table. In our tests, the sub-table contained a large percentage of those interior cells with 3 or more contributors. There is another cell-based extension of the EZS method that we explored but did not report on here; it applies ‘balanced noise’ to non-sensitive cells and ‘uni-directional noise’ to sensitive cells. Which variation is best for a given survey depends on the characteristics of both the microdata and the tables. After testing the application of noise to several quite different surveys, we believe the existence of several variations of EZS noise makes it likely that a given survey can be well protected by EZS noise and will produce data products of good data quality.

In our future research on noise protection, we plan to explore some issues that have a strong policy component such as whether or not some (or all) sensitive cells should be suppressed and if cells that are greatly perturbed should be flagged as such? These questions are two of many that fall under the task of determining how much users can be told about the noise process without compromising data protection. We would like to reveal enough information about the noise process to make it possible for a user to associate usefully narrow uncertainty intervals with cell values. There is also the issue of how to protect more than one magnitude variable from a given survey. When is it a good idea to use the same noise factor for two variables?

We also plan to explore the many complexities that arise when one tries to apply the noise process to an annual survey. Should the same noise factor be used for each (still existing) company? How should noise factors be assigned or modified when there are new companies coming into existence (i.e., company births) and company mergers and acquisitions?

Finally there is the challenge of coordinating the protection of multiple surveys that use overlapping data. If one survey uses one variation of the EZS method, can another survey use another variation or even a different type of protection method such as cell suppression?

Acknowledgements

Thanks to Laura Zayatz of the Disclosure Avoidance research group at the U.S. Census Bureau and many members of the Economic Directorate at the U.S. Census Bureau. We benefited greatly from discussions with members of this group on a variety of aspects on the use of noise to protect tabular data.

References


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

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