An Overview of Uncertainty Creation to Protect Statistical Data

Paul B. Massell
SRD, U.S. Census Bureau, 4600 Silver Hill Road, Washington, D.C. 20233
Paul.B.Massell@census.gov

Abstract[1]

Statistical data releases from statistical agencies (e.g., U.S. Census Bureau) have two basic requirements. These are (1) the release of useful summaries of the contributed data that often are presented for a wide range of geographic levels (e.g., from national to small geographic regions) and (2) the need to protect the contributed data from disclosure, however that is defined in a given situation. We discuss a few protection techniques that satisfy those requirements, albeit in different ways. Many of our examples will involve two methods used to protect (economic) magnitude data tables. These are cell suppression and a simple type of noise that is added to microdata values. Some of our examples will involve data swapping to protect (demographic) categorical data tables. However, some of our ideas are quite general and should be applicable to other protection methods. To ensure that the agency adds a near-optimal level of uncertainty to meet the above requirements, the agency needs to (1) estimate all the major sources of uncertainty about microdata values that derive from the form of the data products (e.g., sampling variances for each cell of a table to be released), (2) estimate any uncertainty-reducing methods a clever data user could use that involve all the data products from the agency (e.g., methods that could be used if microdata are not protected consistently in various data products), (3) estimate the prior data knowledge that the best-informed data users will have about contributed data values (e.g., some users of economic tables probably have a rough idea of the sales values of large corporations based on various public data sources).

Key Words: disclosure avoidance, confidentiality, microdata, noise distributions, uncertainty intervals

1. Introduction

There is a tradeoff between the need for statistical information at a fine level of detail in data products and the confidentiality of the microdata used on which these data products are based. There are many quite different techniques that have been developed to protect the confidentiality of data. It's not surprising that different types of data products often need quite different protection methods, but it is surprising is that for a *given* data product, there are sometimes very different protection methods available. In this paper,

[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.

we try to identify some common features of these disparate methods, because we are seeking to identify the key concepts underlying the goals of statistical disclosure analysis. In each case, we ideally create a certain level of uncertainty in the mind of data users regarding the 'individual level' data while (ideally) minimizing the uncertainty regarding the associated 'statistical level' data. Two major points that we address below are (1) the form of the uncertainty that is created and (2) how the agency releasing the data product can decide on an appropriate level for the uncertainty. Our approach will be to evaluate three techniques (with which the author has some experience) with regard to these goals of disclosure analysis. After some general ideas are formulated using these three techniques, we will try to apply these ideas to a few other techniques.

2. A few protection techniques

2.1 Swapping microdata records based on disclosure risk from associated tables

"Microdata record swapping" is a protection method that has been used to protect the microdata underlying demographic tables generated at the U.S. Census Bureau. These tables are based on Decennial Census data or American Community Survey data. Most of these tables are count tables, although a small percentage are magnitude data tables. The simplest form of this type of swapping is one in which all data items in a record are moved from one geography to the geography of its swapping partner. Two key issues for this type of swapping are (1) the rule for deciding which records should be swapped and (2) the method for finding a swapping partner for a record that needs to be swapped.

Consider a certain class of records that should be swapped. Suppose we release a set of tables with demographic data for households at the block level. A confidentiality problem arises because there are blocks in some regions of the U.S. that have only one household. If a set of tables is published for such a block, a data user can easily acquire data for this household. Some of these data may be sensitive to a household member (e.g., with decennial data, relationships within household and racial identification of family members). For ACS block group level tables (available for a 5-year sample) household income is published but may be sensitive. (It is important to note that the Census Bureau views all variables as sensitive; i.e., all individual data must be protected.)

Focus on the uncertainty aspects of this protection method and assume that we are able to identify all those records that have some disclosure risk. Among these it is natural to define several risk categories ordered by risk level. Those households in a block of size one would likely be in the highest risk category. There might be other records in that category. Generally risk is based on uniqueness of the record, especially uniqueness with respect to variables that would likely be known to data intruders (including snooping neighbors). Sometimes the expression 'swap flagging criteria' is used to denote those criteria which determine the risk level assign to a microdata record. Now assume all records in the microdata file have a risk level assigned, we want to swap all those records in the highest level categories, but perhaps a decreasing percentage in the lower risk categories as risk decreases. A lower percentage might suffice because the chance of a successful data intrusion for these less risky records is tiny. If we are convinced of that, then only a low percentage should be swapped because there is likely to be increasing damage to data quality as the swap percentage increases.

However, even these limited swap goals may be hard to meet in practice. The reason is that certain quantities must be preserved under a swap. The most important one is the household size. Also important is the count of household members 18 years old or older. These counts must be preserved because they are used to determine representation levels of jurisdictions. The advantage of swapping only a small percentage is that a higher level of data quality is maintained. Of course, if geographically close swap partners can be found in (almost) all cases, the statistics of most geographic areas will be unaffected by the swap. For example, if a block group consists of 20 blocks and a number of these blocks have only 1 household, then it might be possible to find swapping partners for each one within the block group. If that were the case, the block group tables would be unaffected by the swap.

Swapping just a small percentage of the low risk records creates uncertainty in the mind of a data intruder. Exploring a large number of high risk data products (e.g., blocks of size one), an intruder will sometimes get confused by what he sees. For example, he may expect to gain racial information about a family based on what he knows --- the parents are white, but the block level table may indicate the parents are black. So he may give up his search for information about the family. By creating a certain level of confusion, when the data intruder actually does come across the true values of this family of interest, in, say, a nearby block, he cannot be certain that the values he sees in this incorrect location are the true ones. This illustrates a general point. There is protection even when the data intruder has access to the true value, but cannot be confident he has the true value. The advantage of a high rate of swapping is that it increases the likelihood that the data intruder will run into confusing data. Confusing data is likely to discourage him from further snooping. It's important to note that the data have this confusing quality only to data users who are trying to identify individual households in the data. To a data user who is interested only in statistical uses of the data, the data will not be confusing.

Above, we discussed a specialized case of swapping. The general idea of swapping can viewed as "record-splitting". To illustrate, let Record 1 be expressed as an ordered list of values of data items: (x1,x2,x3,x4,x5) and Record 2 as (y1,y2,y3,y4,y5). Then select a subset of data items to be swapped --- say x1, x3, and x5. After the swap, the modified Record 1 has the values (y1,x2,y3,x4,y5) and the modified Record 2 has the values (x1,y2,x3,y4,x5). Generally, when dividing data items into two subsets, one subset staying with the record and the other being swapped, we can harm correlations that involve variables from both subsets. (In the specialized type of swapping discussed above, in which geography is the only split-off variable, correlations of variables in regions in which all of a given set of swapped records lies, will be unaffected.) When the swapping is clearly over-protective, it may be possible to use a swapping technique that preserves more variables or specific combinations of variables. This will improve data quality. The optimal implementation of swapping requires knowledge of what makes households vulnerable to data intruders and what data quality aspects are most important to data users.

2.2 Cell Suppression to Protect Economic Magnitude Data in Tables

Let's review the basic facts about cell suppression for magnitude data in tables. In the case of a two-dimensional table, a cell is defined by a "line of work" descriptor for the establishment (the row category) and a geography descriptor for the establishment (the column category). The tables are additive (i.e., there exists a row total and a column total). There are many companies that have two or more establishments. Typically a value is reported for each establishment, and the company value is computed by the agency to be the sum of the establishment values. The agency is required to protect all the establishment values, the company value, and all other values associated with subsets of the full set of establishments.

To protect an establishment value, (e.g., a sales value for some detailed line of work and some detailed geography), we want to create uncertainty about published and suppressed cell values that leads to at least a p% level of uncertainty. There are two common ways to do this: (1) sliding protection--- in which an interval about the true establishment value v is created that is at least 2*(p/100)*v wide and (2) two sided protection--- in which an interval is created that is at least (p/100)*v on each side of v. There are some additional features of the uncertainty interval which are desirable but not required: (a) we would like the true value to be not close to the midpoint of the uncertainty interval much of the time or (b) (even better) for the true value to have a significant likelihood of lying anywhere in the uncertainty interval. Implicit in the above discussion is a notion of a distribution that is associated with the uncertainty interval. Such a distribution could be constructed by exploring the location of the true value in the normalized uncertainty interval for all such intervals for a given set of sensitive cells for some table (perhaps over various cell suppression patterns).

The width of the uncertainty interval depends on the amount of protection flow that has passed through the cell. “Primary protection flow” is based on how much additional protection a cell requires to be fully protected according to the p% rule. “Secondary protection flow” has to do with the amount of flow that the cell contributes to the protection of (other) sensitive cells. If the cell appears in only one table, the uncertainty interval is found by computing the maximum upward and downward flow through the cell take over all flows computed to protect sensitive cells in the table. If a cell appears in two or more tables, these maximal flows are computed from all suppression patterns in any table which contain the given cell.

Example. Table 1 below contains three cells (Pi, i=1,2,3) which require 10 percent protection. There is a solution to the protection flow model found for each P. The cells labeled C are used to provide the needed protection. The “flow” through each cell is labeled with the Pi cell that it helps protect. Some of the P and C cells are included in the solutions for more than one P cell. The solutions are found here by hand, but for a production table the solutions are typically found using an LP model or closely related model. For 2-sided protection, it’s necessary to find 2 flows for each P cell; one in which the needed protection is given a positive sign, the other in which it is given a negative sign. In many cases, the flows for these 2 cases are identical except for sign reversal. In Table 1, we display only the positive flow for each P cell.

Table 1: Initial Protection flows for each of 3 P cells

P1;20; prot=2 (+2 for P1)	C;25; (-2 for P1) (-2 for P2)	P2;50; prot=5 (+5 for P2) (-3 for P3)	P3;30; prot=3; (-3 for P2) (+3 for P3)	125
C:50; (-2 for P1)	C;40; (+2 for P1) (+2 for P2)	C;60 (-5 for P2) (+3 for P3)	C;30 (+3 for P2) (-3 for P3)	180
additional rows…	…	…	…	…

The uncertainty intervals displayed in Table 2 reflect the result of both flows (i.e., positive and negative) for each P cell. The intervals are wide enough so that they could be published without revealing too much about underlying true values. In tables of realistic size, for almost all P cells the uncertainty interval is *not* symmetric about the true value, so a data intruder could not simply assume the mid-point of the uncertainty interval is close to the true cell value. If the agency considers that the level of asymmetry is not high enough to protect some of the P cells, there are ways to introduce more asymmetry into the protection process.

Table 2: Table with Uncertainty Intervals Corresponding to Flows

[18, 22]	[23, 27]	[45, 55]	[27,33]	125
[48, 52]	[38, 42]	[55, 65]	[27,33]	180
additional rows ….	…	…	…	…

2.3. EZS Noise to Protect Economic Magnitude Data

Before discussing EZS noise, it is useful to define two general ways of protecting tabular data.

TABLE LEVEL protection. In table level protection, one forms a preliminary version of the table without any regard to protection, then determines which cells are sensitive (according to some rule), and suppresses or modifies the values of certain cells (e.g., the sensitive cells and often some others). This modification process occurs in stages, (e.g., by protecting one sensitive cell per stage), so that when the final modified table is constructed, all the sensitive cells are protected. The cell suppression method discussed in section 2b was an example of this type of protection.

MICRODATA LEVEL protection. Consider all the microdata values of a magnitude variable whose value forms the cell values in the tables of interest (e.g., in a business table, the variable might be 'sales in dollars for year 2007'). For each such microdata value, we can compute a modified value that will be a new variable in the microdata file. The purpose of the modified value is to protect the true (i.e. unmodified) value. The modified value may be computed before any tables are formed, or after preliminary tables are formed. In any case, before the formation of the final version of the tables there is a single fixed modified value for each true value and all subsequent tables will be tabulated using the fixed modified values.

There are a variety of ways of computing the modified values. They could be computed using a deterministic formula or by using a random number generator. In the case of multi-stage computation, one method could be used for the initial modified values, and other methods for the other stages. For the case of EZS noise, the initial modification is based on a (noise) multiplier nf (called a noise factor) that is drawn from a simple probability distribution. To preserve some statistical properties, it makes sense to require that the distribution be symmetric about 1 and close to 1 so the modification may be viewed as a perturbation (i.e., the modified value equals nf*v and the net perturbation equals (nf-1)*v where (nf-1) is typically small, say, in the interval [-0.5, 0.5] or a subset thereof). The initial modified value is simply (nf)*v where v is the true microdata value for a given microdata record. In EZS random noise, which is the version discussed in reference [2], there is only a single stage of modification and it is based on a draw from such a simple symmetric distribution.

In multi-stage modifications, this initial value may be modified further using cell information from the first version of the table that is formed from the true values or from the stage 1 modified values. In EZS balanced noise, a table is formed from the stage1 modified values. Then a subset of the table cells, called the assignment sub-table, is defined. For each cell in the assignment sub-table with a sufficient number of contributions, an effort is made to decrease the net noise in the cell value by, in sequence, choosing a noise direction for contribution (k+1) that is opposite from the net noise based on the noise contributions 1,...,k. (Usually the sequence is in decreasing order of noise in the contribution). This is a simple greedy algorithm for minimizing the amount of noise. The net noise in each assignment table cell is not necessarily equal to the actual minimum, but often is close to it. Minimizing the net noise for these cells makes sense, because they are usually non-sensitive so there is no reason for them to be have a large distortion.

In the case of EZS random noise for unweighted microdata, the noise distribution is calibrated to the amount of perturbation that is required for protection according to the agency's disclosure protection requirements. If the agency requires a 10 percent perturbation, the density function would be zero from 0.9 to 1.1, (and, say, non-zero on the intervals [0.5, 0.9] and [1.1, 1.5]). This ensures that microdata values are modified (either up or down) by at least 10 percent. The agency might decide not to publish a cell value based on only one contribution (i.e., it might suppress them). If such values are suppressed, a data intruder might be able to recover them**,** but his ability to obtain the unmodified true value is still limited. Assume then that the data intruder is viewing a value from such a cell, and the value is 110. The intruder may suspect that the value has been perturbed by a certain percentage, but he likely won’t know the exact percentage. Should he correctly guess that the minimal percentage that the agency uses is (say 10%) he knows that either v*nf=110 where nf > 1.1 OR nf < 0.9. The corresponding cases for v are v < 110/1.1 (=100) or v > 110/0.9. = 122.2. So even with a good guess about the parameter, the intruder does not gain much information about the true value v. If however, he has some information from a separate source, which suggests that v > 100. Then he can combine this external information with information from the Census table to deduce that v > 122.2. However, this range is wide, and therefore his knowledge is still very rough. Thus, even assuming the data intruder has one key noise parameter and some prior knowledge of the range of v, an agency will likely conclude that v is protected.

3. Comparative analysis of uncertainty creations in various methods

Creation of uncertainty to protect statistical data can take many forms depending on the nature of the data and how protection is defined for that data. For the method of swapping, the goal of uncertainty is to make identification of household records impossible for data intruders who possess partial information about the household and wish to gain more. Precise measurement of disclosure risk and uncertainty are difficult for this type of data. For magnitude data tables, the goal of uncertainty lends itself to more of a mathematical description. We want to modify the cell values so that even a determined data intruder using a variety of mathematical and statistical tools cannot estimate, with confidence, where on the uncertainty interval for a cell value, the true cell value lies. If the data intruder cannot form a good estimate he will be unable to proceed to gain a good estimate of the underlying microdata (e.g., company data). Protection of the latter is the ultimate goal of uncertainty creation.

It would be desirable to add more mathematical and/or statistical precision to the notions of disclosure risk and uncertainty creation. It’s unlikely that data independent protection procedures are sufficient; i.e., it appears that certain features of the data always play a role. With magnitude data tables, there may be some microdata sets which generate tables whose cell suppression patterns lead to uncertainty intervals that are often far from symmetric about the true values. In that case, for the set of cells whose true values cannot be released, it may be acceptable to instead release uncertainty intervals. For other microdata sets, asymmetry must be added.

From the methods discussed in this paper it appears that a significant amount of computation is required to compute risk and to create the proper amount of uncertainty. With swapping, one must do a ‘uniqueness analysis’ on the household microdata of interest before one can estimate the level of swapping required. After swapping, uncertainty can be estimated by computing the probability of success if a data intruder tries to extract information about specific households from the tables. If the level is deemed not adequate, additional swapping can be performed. That is, the process may require iteration.

4. General Remarks about the Use of the Idea of Uncertainty in Disclosure Analysis

Uncertainty analysis is a new approach which tries to identify all the sources of uncertainty that are relevant to the analysis of a problem (ref: Ayyub, et al). If the problem analysis requires a decision be made, it likely will be necessary to make some rough assumptions for each of the sources of uncertainty. The agency may need to make some quantitative assumptions e.g., it might assume that well-informed data users know certain magnitude values within a factor of 2 prior to release of data products. For demographic data, one might assume that people know certain facts about the basic demographic variables of their nearby neighbors. Over time, as additional information and insight is gained, these assumptions can be refined and eventually they may reach the point where traditional probability and statistical models can be used. To make this concrete, consider assumptions that an agency must make about how much informed users know about microdata generally, or for specific records, prior to the upcoming release of new data products by the agency. In this case, by ‘additional information’ we are referring not to the objects being described in the data products (i.e., the sampled units), but the agency’s information about the **users’ prior knowledge** of the microdata underlying the data products.

Our examples so far have been mainly about assumption of users’ prior knowledge of microdata. This is an issue that is probably more important in disclosure analysis than in most other types of statistical analysis carried on by statistical agencies. However, uncertainty analysis does arise in other types of analysis that arise in survey research. The expression ‘total survey error’ refers to the list of all possible sources of error in the survey process that could contribute to error of the released data products. A discussion of some of these sources is very useful to gaining a better idea of how we are applying the notion of uncertainty analysis in this paper. Consider 4 important stages of the conduct of a survey; survey design, data collection, data protection, and data presentation. Survey design is the area which is the most developed mathematically. There are several books on how one should compute sample variances for many types of survey designs. When one can compute a variance based on well-developed probability arguments, we have an example of the most precise type of uncertainty analysis. More typical of uncertainty analysis is the case of response error as a source of error in the data collected. This is hard to model but it is important to try nonetheless, even if one can develop at best a rough model. As discussed above, uncertainty analysis for disclosure analysis, ranges from components that are quite developed mathematically, e.g., protection flows in cell suppression, to the type of disclosure analysis used in swapping, which is much more computational and may require iteration. Even for the cell suppression example, one needs to be more precise: the mathematics is sufficient for uncertainty analysis only in cases when user prior knowledge can be safely assumed to be minimal. In other cases, the mathematical analysis is based on a rough knowledge model. The last of the 4 types of error we are considering here is data presentation, e.g., rounding of cell values in a table, or table design, in which one decides how the rows and columns for a table are defined. These decisions are often made without regard to disclosure issues. However they do create some uncertainty about the values being displayed. The main reason for considering all (or as many as feasible) sources of uncertainty in the data, is that the goal of disclosure analysis is to ensure that an adequate level of uncertainty is associated with each microdata value. If uncertainty analysis of all survey errors shows that uncertainty of all microdata values is adequate (to fully protect such values from disclosure) no additional uncertainty needs to be added via specific disclosure avoidance procedures.

5. Conclusions

Based on the methods discussed in this paper, it appears that a unified approach to disclosure risk and uncertainty creation will be challenging. However, progress along those lines would help in the comparison of methods that are very different in form.

Currently, comparisons are often done using qualitative descriptions. More formalization of ideas might lead to either formulas or algorithms that could be applied to new datasets to give a quick quantitative analysis of the disclosure risk of a dataset and the amount of uncertainty, if any, that must be added to protect it

References

Ayyub, B.M., G.J. Klir, “Uncertainty Modeling and Analysis in Engineering and the Sciences”, Chapman and Hall, 2006.

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

Evans, Timothy; Laura Zayatz, John Slanta (1998), “Using Noise for Disclosure Limitation of Establishment Tabular Data”,J. Official Statistics http://www.jos.nu/Articles/abstract.asp?article=1445333.
Groves, Robert, “Survey Methodology”, Wiley 2004 (discusses total survey error).

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

Massell, Paul; Laura Zayatz, Jeremy Funk, (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, PSD 2006, Proc. Lecture Notes in Comp. Sci. (LNCS) 4302, Springer 2006, ISBN 3-540-49330-1.5.

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

Zadeh, Lofti, (papers on uncertainty in computer science and ordinary communication) http://www.eecs.berkeley.edu/~zadeh/

Zayatz, Laura, “Disclosure Limitation for Census 2000 Tabular Data”, presented at
Joint United Nations ECE/Eurostat work session on statistical data confidentiality (2003)
http://www.unece.org/stats/documents/2003/04/confidentiality/wp.15.e.pdf