A Less Intrusive Variant on Cell Suppression to Protect the Confidentiality of Business Statistics

G. Sande

Sande & Associates, Inc.
10 Regency Park Drive #604, Halifax, Nova Scotia B3S 1P2 email:g.sande@worldnet.att.net

Abstract: Cell suppression is the standard method used to protect the confidentiality of business statistics. It has little effect on highly aggregated statistics. However, it can be very intrusive for highly disaggregated statistics. Some have suggested perturbed values be used instead of withheld values. The use of rounding to variable bases preserves the strengths of cell suppression without the drawbacks of perturbation.

Keywords: Statistical Disclosure Control, Statistical Confidentiality, Business Statistics, Cell Suppression

1 Introduction

Cell suppression is the long standing standard method used to protect the confidentiality of business statistics. It was developed in parallel with the introduction of extensive use of statistical data in business. Statistical data is so useful that institutions have been developed to provide it[1]. These institutions are the statistical agencies. They are given both the privilege of accessing the data and the responsibility of ensuring that the data is useful and can only be used for statistical purposes. In this note we will review the problems that cell suppression poses for the users of the statistical data as well as various methods that have been suggested for addressing these problems. One of the methods is a technique which has not been previously fully explored and will be developed later in this note.

2 Data

The data with which we are concerned is business economic data as found in a typical business survey such as a Census of Manufacturing. Businesses are described by their locations and lines of business.

The location is the common political jurisdiction of nation, region, state, county and locality, which is a generic name for towns and cities. The particular names for the levels of this hierarchical structure may vary. A complication is that metropolitan areas, which are groupings of localities, could serve as an alternate grouping instead of counties. Sometimes a metropolitan area may overlap state boundaries as it is often the case that major cities are located on rivers which also act as state boundaries. A major inconvenience with geographical coding is that the sizes of states may vary widely, with the smaller states being much smaller than the larger cities. A common form of article for magazines is a proposal for a more rational geographical grouping which is based on current economic circumstances rather than historical developments. The protection technique of table redesign is regrouping applied to tables. Business statistics publications which did not follow conventional political boundaries would not be viewed as acceptable so geographical redesign is not practical.

The line of business is described by a Standard Industrial Classification (SIC) which is subject to periodic revision, and even name change. SIC revision is a practical form of table redesign over time. There is less disparity in the activity under the various SIC codes than for the geographical codes but it is still true that the activity under the more active SIC codes will be much greater than that under the less active SIC codes.

Economic theory of the firm suggests that the size of any business will be the accumulation of successive increases or decreases, with the increments being multiplied, so that the sizes become highly skewed. The economic activity of any firm in a given time period will vary about its size. The collection of economic data will be highly skewed but the value of any firm will be centered on its size which may be assumed known to anyone with some degree of interest[2]. Larger businesses are not located at single geographic sites but have multiple locations for their multiple components. The various components are called establishments in statistical practice to allow for branches, subsidiaries and other legal forms. The entire business is usually called an enterprise. An enterprise is a collection of one or more establishments that share their internal economic data and is considered to be the respondent in statistical practice.

The economic statistics with which we are concerned are simple tabulations of the economic activity classified by geography and SIC. The resulting tables are of great interest when they are of industry groups or major industries at the level of the nation or region. When viewed at the level of counties, or even smaller states, and individual industries the tables become rather sparse. A description of Swiss cheese tables is suggestive of some isolated data in the presence of many zero values. Additional classifications are regularly used to study ownership, trading patterns and other attributes of businesses. The publications often have several data fields in each tabulation entry displayed across a page with the entry label listed down the page so that the absence of data is not

1. Antitrust legislation makes the private sharing of business data by business trusts against public policy. Statistical agencies provide public sharing of business data. Business data is so useful that business trusts had been formed before statistical agencies served that function.

2. In many industries there will be either trade association or proprietary databases listing all the production facilities with nominal capacities and technologies implemented. Access to these databases typically requires being part of the industry, signing nondisclosure agreements and paying substantial fees. These conditions are easily met by active businesses but may be difficult for casual observers.

highlighted. When a single data field is organized as a conventional table the sparsity becomes much more evident.

2.1 Data Sensitivity

The data collected from the enterprises is private and is to be considered confidential. The protection of this data is done by aggregating it with similar data so that the details of any respondent are not known. The amount of aggregation required to protect the respondents is embodied in the notion of a nonsensitive aggregation. The various special cases of a single respondent being explicit disclosure, of two respondents being disclosure of each respondent to the other, of a large respondent and a small respondent being little different than just a single respondent and of two large respondents and a small respondent being little different than just two respondents are often listed and would have been recognized when industrial statistics were initially published. These can all be collected into a rule which identifies an aggregation as sensitive if the largest two respondents are more than some percentage, typical values are 75 to 95 percent, of the total of the aggregation. Such a rule would be called an 2 respondent *k%* concentration rule in current terminology. It is a special case of an *n* respondent *k%* concentration rule, commonly called an *n-k%* concentration rule. The technical complication in this is that while a marginal aggregation's total is the sum of the internal values the sensitivity of a marginal aggregation must be determined directly. The size configuration, or even count, of the respondents is variable as an enterprise may have establishments in several of the internal cells. The motivation for *n-k%* concentration rules does not explicitly use the assumption that there is some level of general knowledge about the size of the respondents although it is deals with the same general concerns. Some agencies would apply more than one concentration rule to deal with perceived inadequacies of a single concentration rule. All of these rules were directed as classifying an aggregation as either sensitive or nonsensitive.

The advent of automation lead to further analysis of the properties of sensitivity rules. One objective was to yield a value for the amount, or degree, of sensitivity rather than just the classification result. The most basic requirement was that aggregating two nonsensitive aggregates should result in a nonsensitive aggregate. This property is now called subadditivity. Subadditivty can be traced back to the possible reordering of contributing enterprises under aggregation. Subadditivity and linearity lead to the linear sensitivity rules which apply decreasing weights to the ordered list of respondent contributions to an aggregation. The *n-k%* concentration rules are a special case of linear sensitivity rules where we can identify the property of being sensitive with a positive value for the numerical sensitivity value. If we use the assumption on the general knowledge about enterprise size we are lead to the *p/q* concentration rules. Here the second largest respondent should not be able to use the aggregation total and a *p%* approximate knowledge about the smaller respondents to obtain a better than *q%* approximation to the value of the largest respondent[3]. The *n-k%* concentration rules are recognizable as approximations to *p/q* concentration rules for the values of *n* and *k* or *p* and *q* typically suggested for practical use. Variations on the assumptions lead to corresponding variations in the technical form of the sensitivity rules. The *n-k%* concentration rules are now seen as sensible but obsolete approximations to p/q concentration rules that should no longer be used.

2.2 Data Withholding

When an aggregation is identified as being sensitive it is withheld from publication. Rather than a number there would be a withholding symbol, with *c* for concentrated, *d* for disclosure, *w* for withheld or *x* for crossed out all being in use by various agencies. The withholding will only be applied to the fields which reflect the business operations. The number of enterprises responding to a cell will be always be published as businesses identities are well known. The need for additional, complementary is a common technical description, withholding is readily recognized to avoid determining the withheld value by simple arithmetic. A sensitive aggregate and its complementary withholding are a new *ad hoc* aggregate which have a known, after minor processing, value. This *ad hoc* aggregate would also require sensitivity testing and in cases of high sensitivity there may be several additional complementary withholdings. In a two way tabulation this would have to been done in both directions of the table. The complements would also require protection. Clerical procedures based on completion of the corners of a rectangle are quite effective in manageable sized tables. The rectangles can become more general paths in the presence of empty cells where desired corners might have been. In the presence of general paths and hierarchically structured collections of tables these procedures become clerically unmanageable.

Automation based on following the clerical rules tends to be unsuccessful as with limited rules the implementors of the automation cannot reproduce the common sense of the experienced clerks. Automation based on numerical criteria and mathematical programming is able to reproduce the clerical common sense and deal with the complications of general paths and elaborate able structures.

Complementary suppression in very sparse tables will turn a Swiss cheese table into a display of withholding symbols and a light sprinkling of data. If the publishers of the data did not attempt to publish at such a level of detail they would be subject criticism for failure to fully use the data which is expensive both in terms of respondent effort and statistical agency costs. Instead they are criticized for frustrating the users by withholding so much data. One would expect technical difficulties both in the presentation and in the use of

3. These rules only use the ratio of *p/q* so may be called *c* times improvement rules for *c* having the values of *p/q*. They may also be called *p%* rules where *p* is the *q* of the ratio and the *p* of the ratio is fixed at *100%* for an assumption of positive data or external knowledge of ±100%.

data as the boundaries on what is publishable are approached. The technical difficulties would be there both for the producers in approaching the boundary and for the users in dealing with the resulting technically elaborate publications.

2.3 Older Criticisms of Data Withholding

When respondents to statistical agency publications are asked about their experience with use of data subject to withholding the answers are somewhat equivocal. They want both more withholding in those part of the publications which provide data about their operations and less withholding in those part of the publication which provide information about their competitors. One can only understand this to say that they expect that there will be confidentiality protection and that it will impede some types of analysis. One gains further insight into their needs by looking at the various special requests for *ad hoc* tabulations that are made to supplement the publications. Many of these are for the same type of information as is published but for differing aggregations, such as marketing areas as defined by their criteria. Others will be for further disaggregation under some attribute not used in the regular publication. Those working on the surveys occasionally express a desire to better indicate the uncertainties in the results. This takes a concrete form in various error indicators and even some amount of quality oriented withholding. The same concerns underlie the comments that the confidentiality procedures are very dependent on manipulation of error free values to cause disclosures so that the protection may be overly zealous.

The special request *ad hoc* tabulation may already be well determined by the existing publication. Obtaining that determination requires that one realize that every withheld value is bounded above and below[4]. The trivial bounds are the surrounding aggregates above and zero below. These can be tightened by treating this as an optimization problem and asking what is the minimum and maximum values the desired *ad hoc* tabulation could obtain when the various bounds such as those just given are consistently applied to the whole table. This application of linear programming is often called a confidentiality audit. The result will be a range of possible values which for other than the lowest level cells may have small relative error. This analysis is technically possible, and even fairly easy for those with suitable machinery, but is often pragmatically not available for many users.

The resulting interval response is quite uncomfortable to many users, and even many analysis professionals. At one time applied mathematics included many activities such as statistics, numerical analysis, operations research and even interval analysis. With increasing specialization each of these have gone their separate ways. Now the applied mathematics specialist who would calculate statistical tables to high precision is not a statistician but a numerical analyst. One of the tools that might be used to assess the accuracy of the computation might be interval analysis (Alefeld and Herzberger 1983, Burkill 1924, Moore 1966). In such an application interval analysis is used to assess the internal errors in the computation rather than the effects of errors in the input data. Interval analysis represents a quite different paradigm for dealing with inexact data than that used in statistics. Even the basic terminology is confused between the two specialities. Within statistics, interval data is usually taken to mean data where the assignment of the zero value is arbitrary, such as the zero in either Fahrenheit or Celsius temperature scales. Rather statistics would describe such data as interval valued data or even as symbolic data (Billard and Diday 2003).

2.4 Newer Criticisms of Data Withholding

Duncan and Fienberg (1999) provide some criticisms of cell suppression in a conference paper and then cite their paper as being critical of cell suppression. It is worth returning to their original paper to understand their criticisms.

In their first criticism they fault a table which has the United States as the total and the fifty states and several territories as internal entries for failing to provide the regional total for New England. It is certainly a bother that such a reasonable large scale region total is not available in the publication. We have seen this above as the results of an *ad hoc* tabulation not otherwise included in the publication. In this case it may well be that the New England total is in fact well determined. They appear to simply assume that because some states in New England have been withheld that the New England total will be unavailable. In practice this is a safe assumption even if it would require technical verification. It is more surprising that such a standard regional total is not available so that this is more a criticism of the design of the publication than of the protection mechanism. If their criticism had been more technically oriented at the lack of some localized and specialized aggregation it would have had been recognized as asking for something that a general publication might reasonably not provide. Given the great disparity in size of the states within New England this could also be reasonably considered as a argument for not following the existing political boundaries but rather some more rational grouping of the region. The difficulties with this alternative are readily recognized.

They also offer a criticism that withholding can lead to misleading statistical inferences. They provide a constructed example of how this might arise. Their example seems to be more an instance of the well known Simpson's paradox on how inferences on aggregates can be quite misleading on the subgroups. This is a general property of aggregation of unequally weighted groups that is not unique to cell suppression. Cell suppression is only an issue because the subgroups are so small that disclosure is an issue, all subject to the example being completely constructed to illustrate their point.

Their final criticism is that intervals are not a form of data that they are comfortable with. They have much company but then there are many who find statistical analysis quite

4. Most business data is positive. No business activity is a zero value. Exceptions are items like profits.

uncomfortable. Finding statistical analysis uncomfortable is not usually viewed as an acceptable reason for failing to use statistical techniques when they are appropriate to the problem at hand. The application of statistical analysis to interval valued data would seem to be a fertile source of research topics for research students. As others have noted, such data is increasingly more common. It would also be of interest to those who are only provided confidence interval summaries of statistical analyses.

Their criticisms are used to motivate their development of Markovian Perturbation of tables. This is intended to protect small entries in tables of counts. Cell suppression is directed at protecting highly concentrated cells where there are a small number of dominant respondents. The contents of the cell is used to obtain further information about the dominant respondents. Cell suppression is not normally recommended for tables of counts as it does not match the protection required for counts. In microdata release one is concerned with small groups as identification of small groups may lead to identification of single respondents which then provides access to the remaining fields in the microdata. The traditional simple examples of disclosures in tables of counts show a small marginal count with a single internal cell with all the count. This is a high association table. An alternate description is that the table configuration shows homogeneous values for all contributors to the marginal value. Knowledge that a respondent contributes to the marginal value also means that they have a high probability of having the homogeneous value[5]. Perturbing small cells will not greatly change the homogeneity of the values when the margin is not also small. Protecting against high association, or homogeneous values, is even harder than cell suppression. Markovian Perturbation, like random rounding, does not achieve this protection. Protecting small cells is desirable as there are other disclosure risks and small cells are often marginal values in other tables.

In citing their original comments these authors often summarize the criticisms by saying that cell suppression destroys unrelated information. This cannot be technically true as the withheld cells are related by being components of common marginal values in tables. One can only understand their comment to mean that the additive relationship structure used in cell suppression is different than the log-linear relationship structure which they are used to analyzing.

They also go on to point out that any statistical analysis of a perturbed table should use a statistical description of the perturbation applied. They have been quite consistent in repeating this important point in their referencing of their criticisms of cell suppression.

2.5 Important Properties of Protection Methods

We may also list our criteria to judge protection mechanisms. A statistical agency is expected to protect the data, to provide useful publications and balance these and other issues prudently. We may collect our criteria under those headings.

The protection of the respondents data by a statistical agency is both a statutory and an ethical requirement. It is also a pragmatic requirement to retain respondent cooperation. A basic requirement is that the protection should be seen to be done. The traditional methods of considerable aggregation, random rounding and cell suppression all share the characteristic that it is readily apparent that protection has been applied, even when technical documentation is no longer present. The protection must be correctly applied. The protection mechanism must be able to correctly deal with elaborate table structures. This is just a specific requirement for technical correctness but it has been an ongoing problem for clerical operations and some forms of automation which do not use mathematically correct decomposition techniques. Another specific requirement for technical correctness is that the enterprise structure be correctly dealt with.

The published data should be useful for the users. A traditional form of data is macrodata used in many places including in the System of National Accounts. Macrodata should be as complete and as accurate as the agency is capable of producing. The protection for macrodata is typically its considerable level of aggregation. Macrodata is often used for historical comparisons, with year over year comparisons a standard item for economic analysis and often reported in the business press. In practice this means that any agency introduced fluctuations, beyond sampling variations, will not be acceptable to users. Historical continuity is important to many users who show their concern over events like SIC revisions. The demand for highly disaggregated data, or mildly aggregated data or mesodata[6], has been growing over time as analytical capabilities have grown. The intrusive nature of cell suppression is very obvious in the mesodata setting. The visible and extensive withholding of mesodata has lead to the criticisms of cell suppression and the desire for a replacement protection mechanism. The effects of confidentiality protection need to be documented so that the effects may be accounted for in any analysis. Confidentiality protection represents the deliberate use of uncertainty to protect the respondents. There are other sources of uncertainty in the data. An integration of these various aspects of the total data uncertainty would be preferable to treating each source separately.

The preplanned publication program is an important aspect of the operations of a statistical agency. The ability to service

5. An example might be the reporting of the grades of students in a service course by the students's home faculty. Some faculties may have students who are both ill prepared and unmotivated for the course but are quite happy to achieve a below average but passing grade. A table which indicated that such a faculty had all of its students receiving their desired grade of below average but passing would be a statistical disclosure even if there were many students from that faculty.

6. Highly disaggregated is an awkward construction in the style of a double negative. Mildly aggregated is a direct indication of the intermediate level between macrodata and microdata. Meso is the standard indication for things which are between micro and macro in scale. Mesodata is a usage which is appearing in other contexts for the meaning intended here.

additional *ad hoc* requests is also part of the service provided, and important as an indicator of future directions of user requirements.

3 Current Proposals

The protection of the confidentiality of the respondents data relies on the existence of perturbations to the data. In the sensitivity rules the possible existence of adequate perturbations is the technical test which is codified into a more directly verifiable form. The protection at the table level is also based on perturbations. We will compare four proposals for protection of which one is based on applying a perturbation and reporting the consequences, another is based on finding a suitable perturbation and reporting the consequences and two are based on finding collections of perturbations and reporting differing representations of those collections of perturbations. The underlying mathematics of the three proposals that find perturbations is similar although the presentation and reporting of the resulting perturbations is different. The proposals are: Noise Injection which adds a perturbation to the microdata and treats the resulting microdata set as protected; Controlled Tabular Adjustment which finds a perturbation to all the table entries which will both preserve many entries, particularly macrodata, and protect the sensitive cells; Cell Suppression which finds a collection of perturbations, one for each sensitive entry, where the perturbations are only in the withheld cells and Variable Base Rounding which finds a collection of perturbations, one for each sensitive entry, where the perturbations apply to each entry which has been subject to the rounding to the various bases.

3.1 Noise Injection

Noise Injection (Evans et al. 1996) is a simple but effective procedure. The proposal for the technique is a thorough working through of the notion of using noise injection for the purpose of evaluating the technique. The proposal reaches the conclusion that the technique is not appropriate for all data products of a statistical agency. Each microdata record for an enterprise that needs protection is given a perturbation multiplier, which applies to all the establishments of the enterprise. In the simplest case we would either multiply or divide by 1.1 if the required protection were ten percent. The level of protection would be chosen analytically and subject to randomization.

The important properties show that the presence of the protection is not apparent. It is easily applied to complex tables and deals with enterprise structure. The macrodata will be subject to perturbation and the mesodata will be complete and perturbed. Documentation is possible but not given in the examples. Alternate error sources are not represented as is the current practice. *Ad hoc* tabulations are either unneeded or pose no operational problems. Additional classification variables may pose operational difficulties as some new cells may be unperturbed and sensitive.

3.2 Controlled Tabular Adjustment

Controlled Tabular Adjustment (CTA) (Cox and Dandekar 2003) uses mathematical programming techniques to determine a perturbation which will both protect sensitive data and preserve the macrodata. The perturbed values are published. The proposal can be viewed as a development of Markovian Perturbation for magnitude data. Each sensitive entry is required to be protected by a perturbation but the sign of the perturbation is initially undetermined. The nonsensitive entries will be perturbed as necessary to ensure that tables remain additive. The undetermined signs are determined by a requirement that the weighted amount of perturbation be minimized.

There is a comment in one of the descriptions of CTA (Cox and Kelly 2003) that feasibility can be a problem so the required protection may be relaxed to ensure that a solution can be found. Feasibility can always be ensured by perturbing progressively higher level aggregates. Many CTA examples show perturbed values for margins, or table totals, to balance the specified perturbations of internal entries. Some examples show the perturbed value to be zero which would be implausible for an entry with economic activity. The feasibility comment presumably applies to the situation where the perturbation of some level of aggregation is required to be zero and perturbed values must remain nonzero. The CTA problem for specially structured examples is equivalent to the controlled rounding problem. Controlled rounding problem solutions always exist for simple two-way tables but do not exist in many cases of more general tables.

The CTA examples are based on entry by entry perturbation without any indication of enterprise structure. The simplest example of enterprise structure would be of two adjacent sensitive entries having establishments from the same enterprise. If each were to require a perturbation of 5 a possible solution would be to have one of +5 and the other of -5. If the aggregation of the two entries is also sensitive the required joint perturbation could be as large as 10 but would typically be smaller if the smaller contributors were different in the two entries. If the joint requirement were for a perturbation of 6 the solutions could be for both entries to have the same sign of perturbation and meet the joint requirement. This requires that the remaining entries balance a combined perturbation of 10. Or one entry could have a perturbation of +11 and the other of -5 to meet the requirement. This requires the remaining entries balance the increased perturbation of 11. However the enterprise structure is addressed it will be a requirement which can be expected to have more feasibility problems than when the requirement is not addressed.

The important properties show that the presence of the protection is not apparent. It can be applied to complex tables but there may be difficulties which may be made even more difficult by the enterprise structure. The macrodata will not be perturbed and the mesodata will be complete and may be perturbed. Documentation is possible but not given in the examples. Alternate error sources are not represented as is the

current practice. *Ad hoc* tabulations will largely be unneeded but those that are required may cause feasibility problems. Additional classification variable may cause feasibility problems.

3.3 Cell Suppression

Cell Suppression (Sande 1984) uses mathematical programming to determine a collection of perturbations which will both protect the sensitive data and preserve the macrodata. The presence of perturbations is published as withholding symbols. The required value for the withholding symbol may be different for each sensitive entry being protected. The perturbations are determined to minimize the weighted amount of withholding.

In clerical practice cell suppression has been plagued by errors. Automation has also been problematic as some automation systems have been based on mathematically incorrect decomposition procedures. Correct systems that work with no fuss have been in place for more than 20 years (Robertson 1993). As such they are no longer being described as they need no attention.

The important properties show that the presence of the protection is obvious. It can be applied to complex tables and deals with enterprise structure. The macrodata will not be perturbed and the mesodata will not be perturbed when available but much mesodata may be withheld. The documentation of where the protection is applied is implicit in the protection. Alternate error sources are not represented as is the current practice. *Ad hoc* tabulations can be subsumed under the existing analysis but may be subject to considerable withholding. Alternate classification variables may pose some operational problems and may be subject to considerable withholding.

3.4 Variable Base Rounding

Variable Base Rounding uses mathematical programming to determine a collection of perturbations which will both protect the sensitive data and preserve the macrodata. The perturbed values are published as a range of a rounded value plus or minus the rounding base. Each value may have its own rounding base. The perturbation contained in the ranges may be different for each sensitive entry being protected. The perturbations are determined to minimize the weighted amount of perturbation.

The important properties show that the presence of the protection is obvious. It can be applied to complex tables and deals with enterprise structure. The macrodata will not be perturbed and the mesodata will be complete and many may have a rounding range. The documentation of where the protection is applied is implicit in the protection. Alternate error sources are be represented by the same ranges and are used to provide some of the confidentiality protection. *Ad hoc* tabulation will largely be unneeded and those that are required can be subsumed under the existing analysis. Alternate classification variable may pose some operational difficulty.

4 Protection by Variable Base Rounding

The suggestion of range publication as a variant on cell suppression seems to have been made almost as soon as cell suppression was used as a technique. The use of informative withholding symbols to indicate approximate values long predates automation of cell suppression. This practice seems more associated with subject matter areas like employment where quantitative indicators were judged to be important. The tables would have an *x* as the general withholding symbol but an *a* for an employment range of 0 to 20, *b* for 20 to 50, *c* for 50 to 100 and so on for several more steps. Automation of cell suppression by mathematical programming techniques relies on an internal determination of perturbation ranges. The observation that these internally available ranges would be of interest to external users is quite natural.

The problem is not how to calculate the ranges but rather of how to publish them for use. Listing the two end points of the ranges would double the size of a publication and require other revisions in the publication format. Lowering the bulk of a range publication and making a usable format are a large part of addressing the problem. Assuming that a format is available, the question of what values should be published must also be addressed. The internal ranges tend to be the

Summary Table of Important Properties of Protection Methods

	Noise Injection	Controlled Tabular Adjustment	Cell Suppression	Variable Base Rounding
Protect Data				
Protection Apparent	No	No	Yes	Yes
Complex Tables	Yes	Yes	Yes	Yes
Enterprise Structure	Yes	Problems	Yes	Yes
Useful Publication				
MacroData Preserved	No	Yes	Yes	Yes
MesoData Available	Yes	Yes	Problems	Yes
Documentation of Perturbation	Possible	Possible	Yes	Yes
Alternate Error Sources	No	No	No	Yes
Prudent Operation				
Extendable	Problems	Problems	Yes	Yes

minimal required ranges that are symmetric about the true value. This leads to a midpoint attack in which a user would just average the two end points. The combination of these two practical problems and the need for users to become comfortable with automation for cell suppression caused the extension to range publication being deferred to a later date.

4.1 Formatting of Ranges

When we seek to address the practical problems of range publication there are two hints that can be found in closely related practice. The first hint is that most experimental values in science are listed to as much accuracy as they have followed by an error indication. There is a convention that when there is no explicit error that one can assume the error of common rounding is implied. Such conventions work well with scientific notation, which is floating point numbers in computing practice. It works less well with fixed width columns. This difficulty applies to high precision table making in numerical analysis where the final accurate digit may be underlined and the following digits are the natural result of the computation process even if they may be erroneous. The second hint is the practice of randomized rounding which suggests that rounding can be for overlapping ranges. Random rounding also suggests that tables of multiples of 2, 5 or 10 are readily recognized and the resulting minor loss of additivity of the tables is well tolerated by users. Random rounding is also an example of the generated ranges not being symmetric about the original data and of the use of non-invertible transformations.

Combining these leads to the suggestion of 2 for 2±1, 3 for 3±1, 4 for 4±1 and 1ø for 10±10, 2ø for 20±10, 3ø for 30±10 and so on. We can also separate 100 for 100±1 from 10ø for 100±10. There will be a requirement of a visually pleasing variant on the place holder ø. A small amount of exploration suggests that one would like additional ranges of 2±2, 4±2 and 6±2 as well as 5±5, 10±5 and 15±5. The need for ranges such as 3±2 or 7±5 seems less pressing. The doubling, or the reasonable approximation of 5 for 4 in decimal numbers, of the rounding base seems quite natural. The larger collection of symbols then suggests that alternate fonts or various stylings of a single font be used to indicate whether the rounding base is 1, 2 or 5 with the same styling used to indicate place holding 0s. This is a sufficiently workable solution to the formatting problem that the question of what values should be represented can also be addressed.

4.2 Values for Ranges

The obvious problem with the internally used symmetric ranges is that they suggest using the midpoint as an estimate. An immediate suggestion would be to double the width and randomly place the extra width on one or the other end of the existing range. Further consideration shows we need to ask whether doubling is the right extension and how one does the randomization as well as how these would interact with rounding. The existing ranges are specified by a midpoint, which is the value of the entry, and a half width, which is the numerical value resulting from the sensitivity rule. The required protection range is the value plus or minus the required protection. This matches the notion of a rounded value plus or minus a rounding base. The first step would be to determine a proposed rounding base by finding the next eligible rounding base equal to or exceeding the required protection. For example, a required protection of 35 becomes a rounding base of 50. Then the entry value would be rounded to the proposed rounding base. Often the required protection range will be enclosed in the range of the rounded value plus or minus the rounding base. However the rounded value may be near one end of the required protection range with one of the rounded end points near the other end and the other rounded end point some distance from the required range. The rounded range will only partially overlap the required range. A required range of 44±8, or 36 to 52, would becomes 40±10, or 30 to 50, which is not what was hoped for. We have just learned of a source of randomization as the value will be randomly placed with respect to the rounded decimal numbers. This would be called an *experiment of nature* with the randomness in the value which will be unknown once it is protected. The solution to the problem of the partial overlap is to increase the rounding base until the rounded range encloses the required protection. The above example then becomes 40±20, or 20 to 60, which does enclose the required protection.

We can repeat this for all the entries that are sensitive and those that are not sensitive with some amount of source of error to be reported. A source of error from the data might be sampling error, temporal allocation error due to reporting reference period or any other subject matter or statistical issue. One would expect that the absolute amount of error would increase with aggregation while the relative error would decrease. This is the case with sampling error. A source of error from the publication might be the common practice of reporting macrodata in large units to lower the number of digits represented. If no error source is available the result will be cell suppression with an explicit range provided for each withheld entry.

We must also address the issue of whether just enclosing the required protection range by a rounded protection range is

Range Formatting Table

Base 1		Base 2		Base 5	
Light Styling (Italics)		Medium Styling (Bold Italics)		Heavy Styling (Extra Bold Italics)	
Symbol	Range	Symbol	Range	Symbol	Range
1	0 - 2				
2	1 - 3	***2***	0 - 4		
3	2 - 4				
4	3 - 5	***4***	2 - 6		
5	4 - 6			***5***	0-10
6	5 - 7	***6***	4 - 8		
7	6 - 8				
8	7 - 9	***8***	6 -10		
9	8 -10				
10	9 -11	***10***	8 -12	***10***	10-15

adequate. If the rounded protection range is a good approximation to the required protection range then a midpoint attack will be effective. Even when the rounded protection range width is doubled a midpoint attack will only just be at one of the boundaries of the required protection range. We would judge this inadequate. If we ask where the required protection range might be relative to a double width provided protection range we see that the midpoint of the required protection range can only be in the central single width, and not in either extreme half width of the provided protection range. Or that the midpoint of the provided protection range will always fall somewhere in the required protection range. A more acceptable situation is obtained if we require the provided protection range to be at least four time the width of the required protection range. Now the midpoint attack will be outside the required protection range for at least 2/3 of the possible positions. The midpoint will be no better than either the one quarter or three quarter point.

4.3 Calculation of Expanded Ranges

We have been able to specify ranges for sensitive entries and for many other entries. We may be fortunate and these ranges may allow adequate perturbations to protect the sensitive entries. Such a fortunate configuration may also arise in cell suppression when the withheld entries provide adequate mutual protection. However we may be required to make some ranges wider to achieve the desired protection in the same way that we may require complementary suppression in cell suppression.

The basic mathematical formulation of the several perturbation methods has the same underlying form. The side conditions and reporting modes for each method are distinct. Each entry in the publication will have a value v_i. Some entries will be lowest level entries which are not aggregations. Other entries will be structural aggregates and may either have a single combination rule, such as a row or column margin, or may have alternate combination rules, such as a table total obtained by summing over both rows and columns. Still others will be *ad hoc* aggregations of sensitive entries, and perhaps even with some nonsensitive entries, which are themselves sensitive.

The relationships of the various v_i may be summarized as $M\ v = 0$ where any row of M will have a single -1, several $+1$s and many 0s. Each v_i may have a perturbation x_i and $M\ (v + x) = 0$ to indicate that the perturbed values satisfy the same relationships. The perturbations will be subject to constraints of the form

$$(1 - p)\ v_i \le v_i + x_i \le (1 + p)\ v_i$$

where p the fraction representing the external knowledge. Common values would be either $1/2$ or 1 with a corresponding value chosen for q. The perturbations for sensitive entries may be further restricted with the values given by the sensitivity rule. The technical details would vary for each of the perturbation methods and the solution technique used for that method.

An objective function is used to choose between the many possible perturbations. In cell suppression and controlled tabular adjustment the terms of the objective function would be $c_i \mid x_i \mid$ for some c_i. Common choices have been 1, v_i or $\log(v_i)$ to minimize the count of perturbations, the value of the perturbations or the entropy of the perturbations. Some use $\partial(x_i)$, which is 1 for nonzero x_i and zero otherwise, rather than x_i to have an integer optimization problem rather than a continuous optimization problem. A range publication would require a more elaborate objective to allow the perturbation to take any value within the permitted ranges at no cost. For a permitted range of $[\ l_i, u_i\ ]$ the objective coefficients would be $c_i \mid x_i - l_i \mid$ for $x_i < l_i$, 0 for $l_i \le x_i \le u_i$ and $c_i \mid x_i - u_i \mid$ for $x_i > u_i$. In mathematical programming, absolute values are often dealt with by separating a variable into a positive portion and a negative portion to use existing algorithms. In l_1 fitting this adaptation is often subsumed into the algorithm to avoid the doubling of the number of variables. This three segment objective can be viewed as the sum of two absolute values for two variables or as three variables with separate coefficients. Specialized piecewise linear algorithms would require only a single variable.

A straightforward adaptation of the greedy sequential heuristic algorithm used in the ACSSuprs (Sande 1999) cell suppression program calculated the ranges. The ACSSuite has an l_1 fitting code so the number of variables was double what it would have been for cell suppression. The adaptation was to update the permitted ranges sequentially rather than updating the objective coefficients as is done for cell suppression. An extension of ACSAudit allowed the ranges to be specified by a rounded value and rounding base rather than a lower and an upper bound. The audited ranges may be contained within the published rounded ranges as their end points are not simple rounded values. The audited ranges will round up to the published rounded ranges.

4.4 Examples and Evaluations

The examples are from one of two datasets used to illustrate the USBC program USBCSUP. Both datasets provide similar illustrations. They are heavily perturbed datasets typical of business data. The coding structure mimics a Census of Manufacturing. The data is both highly skewed and sparse. The availability of such datasets is very helpful.

The low level entries were assigned an error source of half the square root of the value. The error source for the standard macrodata was set to zero. The sensitive entries were assigned protection intervals as described above. Protection consisted of reporting the low level entries with variable base rounded values which were either their error source or the expanded ranges required to protect the sensitive entries. In some cases sensitive entries had their ranges expanded to protect other sensitive entries.

The number of expanded ranges was noticeably less than the number of complements under cell suppression. This matched earlier experience with range publication that had not been fully developed to include formatting for presentation

Variable Base Rounding - High Level Table

	G0	GA	GB	GC	GD
S0	53238166	10842153	15217498	13664118	13514397
S10	2158602	568***500***	60*1000*	600000	388***500***
S11	1610537	506***000***	508***500***	30***4000***	292***000***
S12	138655	*20000*	1***00000***	*20000*	3***000***
S13	4386103	103*9000*	116*0000*	134*3000*	84*5000*
S14	2590986	54*9000*	74*2000*	742***000***	558***000***
S15	4886779	832***500***	129***8000***	98*2000*	177*4000*
S16	917839	165***000***	198***000***	25***4000***	301***000***
S17	1389851	137***000***	57*5000*	35*9000*	318***500***
S18	4384015	72*3000*	112*6000*	109*5000*	144*0000*
S19	1440783	269***000***	448***500***	366***500***	356***500***
S20	6697230	17*40000*	173*2000*	196*1000*	12*60000*
S21	1149716	161***000***	47*4000*	297***000***	218***000***
S22	6237173	145*2000*	156*1000*	171*2000*	151***2000***
S23	526827	61***600***	255***500***	105***500***	104***000***
S24	2419729	71*1000*	83*5000*	577***500***	296***500***
S25	1542605	197***000***	452***000***	351***000***	54*3000*
S26	3418695	48*3000*	118*5000*	97*8000*	77***2000***
S27	3240725	433***500***	653***500***	642***000***	151*2000*
S28	187171	27*000*	58***600***	55***400***	46***200***
S29	3914145	761***000***	125*9000*	91*8000*	97*6000*

Cell Suppression - High Level Table

	G0	GA	GB	GC	GD
S0	53238166	10842153	15217498	13664118	13514397
S10	2158602	568502	601373	600255	388472
S11	1610537	506067	508382	304095	291993
S12	138655	20132	95041	20712	2770
S13	4386103	1038577	1159725	1343155	844646
S14	2590986	549246	741805	742004	557931
S15	4886779	832528	1298531	981955	1773765
S16	917839	165213	198060	253493	301073
S17	1389851	137235	574879	359212	318525
S18	4384015	722822	1125837	1095206	1440150
S19	1440783	269211	448605	366313	356654
S20	6697230	1745344	1731906	1961119	1258861
S21	1149716	160768	473666	297038	218244
S22	6237173	1452303	1561067	1712339	1511464
S23	526827	61632	255520	105549	104126
S24	2419729	711080	834892	577430	296327
S25	1542605	196768	451939	351193	542705
S26	3418695	483312	1185391	977601	772391
S27	3240725	433369	653500	641901	1511955
S28	187171	27005	58624	55430	46112
S29	3914145	761039	1258755	918118	976233

High level table with unrounded macrodata on both margins. Internal entries are variable base rounded. All data is present. No data is withheld under cell suppression.

Variable Base Rounding - Middle Level Table

	G0	GA	GB	GC	GD
S20	6697230	17*40000*	173*2000*	196*1000*	12*60000*
S201	145179	2*8000*	5*2000*	37***200***	27***800***
S202	40034	1***5000***	*10000*	6*900*	*7000*
S203	1949510	7***00000***	5*40000*	2*40000*	48***5000***
S204	584949	6***0000***	7***5000***	5***0000***	4***00000***
S205	478730	10*2000*	225***000***	73***000***	78***600***
S206	106609	1*1000*	57***400***	2***7600***	10*300*
S207	3392219	84***0000***	*772000*	152*7000*	253***000***

Cell Suppression - Middle Level Table

	G0	GA	GB	GC	GD
S20	6697230	1745344	1731906	1961119	1258861
S201	145179	27999	x	x	x
S202	40034	x	x	x	x
S203	1949510	690303	535879	237645	485683
S204	584949	59624	77020	51532	396773
S205	478730	102079	224912	73193	78546
S206	106609	x	x	27529	10335
S207	3392219	839780	772120	1527246	253073

Middle level table with unrounded macrodata on one margin. The other margin and internal entries are variable base rounded. All data is present. Some data is withheld under cell suppression.

Variable Base Rounding - Low Level Table

	G0	G01	G02	G03	G04
S27	3240725	*50000*	4*500*	46***000***	11*900*
S271	71405	*100*	*10*	6***20***	
S272	984914	*6000*	***200***	5*400*	*10*
S273	249223	***500***	***50***	1***500***	***400***
S274	936111	21***000***	***400***	9*200*	*7600*
S275	17166	***50***	*10*		
S276	981906	18*700*	39***50***	29***400***	39***50***

Cell Suppression - Low Level Table

	G0	G01	G02	G03	G04
S27	3240725	46647	x	46018	11919
S271	71405	109	x	x	
S272	984914	6385	x	5343	x
S273	249223	438	x	x	x
S274	936111	20935	x	9241	x
S275	17166	52	x		
S276	981906	18728	x	29374	3953

Low level table with unrounded macrodata on one margin. The other margin and internal entries are variable base rounded. Some data is not present. Much data is withheld under cell suppression.

(Kirkendall and Sande 1998). The objective function used can be varied to lower the number of ranges expanded, the amount by which the ranges are expanded or the entropy of the range expansions in the much the same way that the choice of complementary suppressions can be influenced in cell suppression. The publication ranges for sensitive entries would be the required protection ranges described above with possible expansion. The expansion could either have an objective function cost like the nonsensitive entries or be have a zero objective function cost. A nonzero cost represents an attempt at preserving information about the sensitive entries while a zero cost allows them to be used to provide protection of other sensitive entries without restriction much as with cell suppression.

Some CTA developers (Russell and Kelly 2003) have evaluated the utility of their method by calculating the product moment correlations between both the published perturbed values with the sensitive values and the published perturbed values with the nonsensitive values. The calculated correlation values are very high. A similar evaluation of several objective function variants of the variable base rounded method produced product moment correlations above 0.99 which are also very high.

Side by side examples of tables with variable base rounding and cell suppression are shown. The tables are at high, medium and low levels of aggregation. The high level table has industry groups[7] and geographical regions as classifications. It is a complete table that requires no withholding. The middle level table has major industries and geographical regions as classification variables. It is a complete table that requires some withholding. The low level table has major industries and states as classifications. Only a portion of the states are shown. It is not a complete table and requires moderate withholding as one of the states is a small state. Even lower level tables of industries would be sparse and show extensive withholding. The example datasets do not have industry level coding.

4.5 Other Considerations

At the end of this we are left with variable base rounded tables with complex stylings that do not add up to their marginal totals. The question of why all this is needed will surely be raised many times.

One alternative suggestion may be to just have the analysts use a *safe site* for analysis rather than having the statistical agency produce *safe data*. A response is implicit in the discussion above in which some of the respondents were also analysts seeking more information about their competitors. Such proprietary analysis would not be able to meet the confidentiality requirements for a researcher using a safe site. A safe site may be appropriate for an academic researcher who is not interested in specific respondents. Travel to and ongoing access to a safe site may be sufficiently awkward that only a few highly motivated and geographically advantaged researchers would use such a facility. The safe site only delays the issue of reporting the results in a nondisclosing manner.

Another alternative suggestion might be to use the enhanced presentation to indicate ranges about central values which are not multiples of the rounding base. This would both provide further enhanced content and address the problem of table not adding to their margins. The extreme, and obviously so mistaken that it can only be a rhetorical suggestion intended to be rejected, would be to have the true value with a rounded range indicated. This is just the midpoint attack made very easy, and correct, for the data users. The question is how to separate the midpoint rounding from the range representation. The current suggestion will be too complex for many and the suggestion that one try to encode a second attribute into the styling would seem to be even more complex. The application of confidentiality procedures will no longer be readily apparent, and would be even less apparent after any user activity which did not preserve the styling.

A more plausible suggestion to deal with tables not adding to their margins would be to have the marginal entries be the sum of the internal entries which are multiples of their rounding base. This might be plausible for the midpoints but is inadequate for the range indication. There would be the technical problem of controlling the proposed value to actually be within the intended range. The problem of representing both the value and the rounding base with the range representation is present in the margins. Recall that in real publications with deeply hierarchical classifications there are often more marginal entries than lowest level entries so simple examples can be misleading. Non-adding tables are very common as conventional rounding leads to small errors which are tolerated. The errors from variable base rounding would also be small even though larger than from conventional rounding. The existing practices of cell suppression and random rounding also lead to non-adding tables.

5 Conclusion

Cell suppression is the long standing standard method used to protect the confidentiality of business statistics. The intrusiveness of cell suppression in withholding mesodata is very evident. It is not reasonable to expect that publications which are very close to the frontier of what is publishable without disclosing respondent data will be as easy to use as tables of macrodata. However it is reasonable to expect that attention be paid to usability and that the intrusiveness of the existing practice of cell suppression be lowered if possible. An important empirical question is whether the types of analysis attempted with mesodata are the same as those for macrodata. Informal experience with *ad hoc* requests for tabulations of business data suggest that concerns differ at differing levels of detail. Criticisms of cell suppression that do not address the differing analysis needs would be misplaced.

It is suggested that there should be a balance between usefulness of the publications and disclosure risk. Statistical agencies make the balance strongly favour respondent

7. Industry groups are two digit SIC codes. Major industries are three digit SIC codes. Industries are four digit SIC codes.

protection for statutory, ethical and pragmatic reasons. One of the important parts of protecting confidentiality is being seen to do it in a competent and professional manner. Cell suppression has been prone to technical failure which is unfortunate for those statistical agencies allowing such failures. Any replacement would be subject to technical scrutiny as a result of the technical failures in implementing cell suppression. Cell suppression is very apparent as a protection mechanism so the statistical agencies will be reluctant to abandon the technique. The absence of apparent confidentiality protection may appear to be no confidentiality protection.

The single perturbation techniques of Noise Injection and Controlled Tabular Adjustment do not provide an apparent indication of confidentiality protection. They would require additional publication space to indicate the statistical characterization of the perturbation in already bulky publications. The suggestion that CTA might take the step of not providing the full confidentiality protection claimed is characterized as controversial by those who make it (Cox and Kelly 2003). An inability to fully deal with enterprise structure would be recognized by respondents whose industries have detailed public knowledge of the enterprise structure. The extension of single perturbation techniques to allow for *ad hoc* tabulation requests poses additional feasibility problems even if the volume of such requests is lowered by the presence of complete mesodata. Noise Injection has the problem that it perturbs macrodata so that even its proposers suggest that it is not suitable for all statistical products. Within the single perturbation methods there is an exchange of the controlling of perturbations in macrodata for the introduction of the difficulty in finding feasible solutions.

The multiple perturbation techniques of Cell Suppression and Variable Base Rounding provide apparent indications of confidentiality protection. The indications are an immediate source of user difficulty. Neither withheld values nor intervals are conventional numbers. Those who are only comfortable with numbers will be unhappy. The rounded midpoints of the intervals provide plausible values for those who want indicative results rather than full accuracy. The midpoint of a margin will not be the sum of the midpoints of the internal entries and this will make some users unhappy. Making a table additive is a standard statistical problem that is often solved with iterative proportional fitting. In multiple perturbation methods there is an exchange of the analytical ease of finding solutions for the difficulty in reporting and using those solutions even though they provide good documentation of the ambiguity present. The use of formatted midpoints of ranges addresses the immediate problem of dealing with withholding symbols. Explicit ranges allow refined representation of publishable information and better use of other error sources. The problem of dealing with ranges may be the price for approaching the boundary of disclosing respondents data. It is also an opportunity for research into statistical methods to deal with an increasingly common style of data.

6 References

Alefeld, G., and J. Herzberger, (1983) “Introduction to Interval Computations”, Academic Press, New York.

Billard, L., and E. Diday, (2003) “From the Knowledge of Data to the Statistics of Knowledge: Symbolic Data Analysis”, Journal of the American Statistical Association, 98, pp 470-487.

Burkill, J. C., (1924) “Functions of Intervals”, Proceedings of the London Mathematical Society.

Cox, Lawrence H., and R. A. Dandekar, (2003) “A New Disclosure Limitation Method for Tabular Data that Preserves Data Accuracy and Ease of Use.” Proceedings of the 2002 FCSM Statistical Policy Seminar, Federal Committee on Statistical Methodology, Washington, DC: U.S. Office of Management and Budget.

Cox, Lawrence H., and James P. Kelly, (2003) “Balancing Data Quality And Confidentiality For Tabular Data”, Joint ECE/Eurostat work session on statistical data confidentiality, Luxembourg.

Duncan, G. T., and S. E. Fienberg, (1999) “Obtaining Information While Preserving Privacy: A Markov Perturbation Method for Tabular Data”, Statistical Data Protection (SDP’98 Lisbon), pp. 351-362, Eurostat, Luxembourg.

Evans, T., L. Zayatz and J. Slanta, (1996) “Using Noise for Disclosure Limitation of Establishment Tabular Data”, U. S. Bureau of the Census.

Kirkendall, Nancy and Gordon Sande, (1998) “Comparison of Systems Implementing Automated Cell Suppression for Economic Statistics”, Journal of Official Statistics, 14, pp. 513-535.

Moore, R. E., (1966) “Interval Analysis”, Prentice-Hall, Englewood Cliffs, N.J.

Robertson, D., (1993) “Cell Suppression at Statistics Canada,” Proceedings of the Bureau of the Census 1993 Annual Research Conference, pp. 107–131, Bureau of the Census, Washington, D.C.

Russell, J. Neil, and James P. Kelly, (2003) “Evaluating the Bureau of Transportation Statistics’ Prototype Statistical Disclosure Limitation Software for Complex Tabular Data”, 2003 Joint Statistical Meetings.

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

Sande, G., (1999) “Structure of the ACS Automated Cell Suppression System”, Statistical Data Confidentiality, Proceedings of the Joint Eurostat / UN-ECE Work Session on Statistical Confidentiality, pp 105-121, ISBN 92-828-7747-7.