

Comparing Statistical Disclosure Control Methods for Tables: Identifying the Key Factors

Paul B. Massell
Statistical Research Division
Room 3209-4, U.S. Census Bureau, Washington, D.C. 20233
paul.b.massell@census.gov

Key Words: statistical disclosure control, confidentiality, cell suppression, cell perturbation

Abstract

In recent years there has been much research on statistical control methods for tables. In some cases, new developments in mathematical programming and related operations research techniques have led to great speedups in the implementation of existing methods (e.g., cell suppression). In other cases, there have been new approaches to the method itself (e.g., cell perturbation and cell rounding). The result is that there are likely to be two or more methods that can be applied to any set of tables that must undergo disclosure control. How does an agency or statistical organization determine which method is best to use? In this paper, we try to identify the key factors in that decision and discuss how to apply them to a given method.

1. Introduction¹

When a statistical office has collected survey or census data and plans to release some of this data in the form of tables, there are a number of basic confidentiality questions that must be addressed. First, there must be some assessment of whether there is a confidentiality issue at all for the tables being considered for release. If the variables that define the tables (the spanning variables and, for magnitude tables, the magnitude; i.e., response, variables) are not considered to be sensitive, confidentiality probably is not an issue. If the data come from a survey, and the sampling fraction for the entire sample is quite small, the sampling process itself may provide enough protection for the data to obviate the need for application of a separate statistical disclosure control method.

For situations in which it is determined that some statistical disclosure control (SDC) method needs to be applied, there are still a number of basic questions that need to be answered. Should the tabular data be protected by modifying the underlying microdata followed by creating tables in the usual way from the (modified) microdata? Or should one modify the tabular data (i.e., the cell values) directly without using the microdata or perhaps using the microdata but not modifying it? Sometimes the expressions “modifying at the microdata level” and “modifying at the tabular level” are used to describe these two approaches. This decision requires information about individual tables and the connections (i.e., linkages) among the full set of tables being considered for release. If it appears that linkages are the source of the confidentiality problem, one could consider reducing the set of tables being released in a way that reduces linkages. In addition, one could redesign some of the tables; this involves regrouping the categories for some of the spanning variables.

After considering the disclosure implications of table design and table linkages, a preliminary decision needs to be made about which set of tables to release and the design of each one.

If a decision is made to protect the data at the microdata level, one must decide which currently available method is to be used. Similar considerations arise if a decision is made to protect the data at the tabular level. Deciding which SDC method to use depends on a variety of factors, including the type of uncertainty that one wishes to create. This in turn depends on how much the statistical office knows about how the tables will be used by the users and to what extent the statistical office (SO) can accommodate the needs of all of its table users. Finally, depending on the SDC method chosen, there may be a need to apply an additional program, often called an audit program, that determines how well the SDC program that was used to create uncertainty met its specific quantitative goals.

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

In this paper, we will discuss some of the issues involved in these decisions. Occasionally, one will be able to say that one method is better than another. More

frequently, there are tradeoffs involved in the decisions, and the SO must decide how to weight each of the factors. Thus there are many decisions that need to be made when protecting the confidentiality of tables. To make an informed decision, the SO should try to keep current on SDC methods and be able to calculate at least rough estimates of each quantity involved in the various minor decisions that support the major decisions.

A note about some basic terminology is in order. The topic of disclosure control for statistical data is often referred to as “statistical disclosure control” even though a number of the algorithms discussed in this field are deterministic ; i.e. they have no stochastic or random aspect. What justifies the use of the word ‘statistical’ is that the goal of these methods is to create uncertainty; enough to provide an adequate amount of confidentiality protection. In some sense, this is the reverse of most types of statistical analysis in which the goal is to reduce uncertainty. In addition, statistics is involved in determining the role of the survey design in the protection of the data.

2. Deciding if tables being considered for release need confidentiality protection

The role of sensitive variables and the sampling fraction

Let us consider the case of a count table for census demographic data. Suppose we have a 3D table that has a variable that is often considered to be sensitive, e.g., income. Suppose the table is AGE by RACE by INCOME where age and income each have 10 categories ; and the underlying geography is a small city. Suppose there is only 1 person with a certain age-race combination in the population. Then the table determines his/her income category. Since age and race are often easy to determine visually (at least approximately) this would be an example of a disclosure (of income) for that person.

Suppose now we have the same table for a large city in which census data determines that each cell has a count of at least 5. Suppose we select a sample using simple random sampling with a sampling fraction of f ; where f is less than, say, 0.5 . Then a count of 1 in a given cell says only that one of 5 or more people associated with that cell has an income in the displayed category. The table user has significant uncertainty about which of the five or more people who “fall” in this cell has the given income; this uncertainty would normally be considered sufficient to eliminate any disclosure risk. Here we are assuming that the 4 people who are not in the sample do not collude; i.e., act together to determine who of the 5

people was in the sample. We make similar non-collusion assumptions below.

Notice that the above argument easily generalizes to more complex sample designs in which the sampling fraction may vary from cell to cell. As long as each cell count is greater than one, the sampling process provides sufficient protection against disclosure.

Let us now give similar examples for economic data. Suppose a single table is to be released where rows represent a type of apparel sold (shoes, hats, etc.) and columns represent major cities in Ohio. If the cell value is the count of stores in the given city selling the given type of apparel, this is no disclosure issue because counts of stores involving only the type of business and the location are considered to be public information. If however, like the demographic example given above, we were to create a table in which categorized ‘sales’ were one of the spanning variables, the disclosure implications are not so simple. If the categories were chosen to be quite narrow, a table user might be able to determine a certain store’s sales value within, say, 10%. This might well be considered to be a disclosure. Here we are implicitly describing data from a census since, the cell counts represent all such stores.

Usually a quantitative variable such as ‘sales’ would not be used as a spanning variable for the table, but as a magnitude variable. In this case, it can be treated as a real variable; i.e., it does not need to be categorized. When ‘sales’ is treated as a magnitude variable, the cell value no longer represents a count, rather it represents the total sales for a given year by all stores that fall in the cell. In this case, the SO needs to determine if the information provided by the table is sufficient for a clever table user to get a good estimate of the ‘sales’ of some establishment. This is an example of the magnitude data disclosure problem that frequently arises with economic data.

Suppose now all stores in the sampling frame (apparel stores in major cities in Ohio) are sampled with a small sampling fraction f , say 1/10. Suppose the unweighted counts for each cell are released. Then if there are 10 shoe stores in Toledo, Ohio, and the sample includes only 1 or 2 of them, there will probably be no harm in releasing the total sales for those 1 or 2 stores if those stores have sales that are in the midrange of figures for the full set of 10 stores. If however, there is one store that dominates the local market for shoes, many table users would probably realize that the sales figure corresponds to that store. If store 1 has sales 50 times those of store 2, publishing the sum of sales for those two stores would also present a disclosure problem.

Thus, for any data with a highly skewed distribution, sampling may not provide sufficient protection against disclosure. Actually, for such distributions, sampling designs are often used that make certain that the dominant establishments are included in the sample. Thus the survey is really a census for these dominant establishments and a sample survey for the less dominant ones. This is in contrast to the situation for most demographic count tables, in which each sampling unit for a given cell is sampled at the same rate and therefore has the same weight.

These simple examples illustrate certain common situations. Sometimes publication of a table presents no disclosure risk because of the variables involved. Under certain conditions (e.g., all stores in the sampling frame for each cell have values that are the same order of magnitude) sampling will provide adequate protection. Sometimes the protection can be predicted prior to data collection. Other times, various calculations must be done with the collected data to determine the disclosure potential of each cell. Protection provided by sampling is discussed in ref: WdW2, p.144-150, and Greenberg (1999; informal note).

3. Deciding what types of uncertainties are acceptable to the SO and to users

In recent years, a number of new methods for tabular SDC have been developed. When these methods are considered along with the older methods, there are likely to be at least two methods which can be applied to any table or set of tables that requires protection against disclosure. There are various factors that need to be considered in deciding which method to use.

A. SO tradition

In any SO that has been releasing tables for a number of years, it is likely that disclosure methods are treated as part of 'corporate knowledge' or as 'the corporate way of doing things.' There are some obvious practical advantages to continue using the same methods. The software already exists and people at the SO are familiar with preparing the data input files, running the programs, and examining the output. Thus, if there are no major changes in the size and structure of tables that must undergo disclosure processing, and no changes in SO disclosure policies or protection thresholds, continued use of existing methods may well be preferred by the SO.

B. Satisfaction of confidentiality requirements

The type and amount of uncertainty that is created must be easily understood, at least approximately, by all the persons at the SO with some responsibility to assess

adequacy of confidentiality protection. If the SO has a disclosure review board, the board members would be included in the group of such persons. Any new SDC method that is proposed for providing protection at the SO would typically be explained by a disclosure technical expert using mathematical and statistical arguments and computational evidence. Some subset of the group of persons mentioned above, would then decide if the new method meets the confidentiality requirements of the SO.

C. Software acquisition and processing.

It must be possible to acquire software for performing the proposed SDC method or to develop the software within the SO in a reasonable length of time. The processing speed for the given set of tables must meet the SO's time constraints. Thus, certain methods may be appropriate for small tables, and other methods for larger tables.

D. Usefulness to table users

The tables that are produced must be in a form that is useful to most table users. The table users are the customers of the SO; thus if the tables are not useful to the users because of the way uncertainty is incorporated, the SO should be aware of this. The SO should perhaps encourage feedback from the table users. Of course, producing tables that meet the needs of all users may not be possible. Usefulness depends on various aspects of tabular uncertainty creation.

i. Is the table complete? We say a table is incomplete if there are cells that do not contain a value but instead are blank or contain a special symbol indicating that the cell value was withheld for disclosure reasons.

ii. Is it clear which cells are the 'true' values and which ones have undergone some sort of adjustment or perturbation. By a 'true value' we mean a value that would be published if disclosure were not a consideration.

iii. For those cells that have been perturbed, is it clear what the maximum amount of perturbation is? In particular, can the user easily determine an interval in which the true cell value lies?, i.e., is the interval given or can it be easily calculated?

Let's evaluate some SDC methods with respect to these criteria.

Cell suppression leaves the suppressed cells blank or with a symbol that indicates withholding for disclosure reasons. One might use the term 'undeclared perturbation (UP)' to indicate perturbation in which the cells that have been perturbed are not designated.

‘Perturbation with symbols (PS)’ refers to perturbation in which a symbol is placed next to any value that has been perturbed. ‘Perturbation with intervals (PI)’, refers to perturbation in which a value is given with an uncertainty interval, possibly in the form $v \{+/-\} d$.

Table Property	Cell Supp.	UP	PS	PI
Complete table	No	Yes	Yes	Yes
Modified Cells are Designated	Yes	No	Yes	Yes
Uncertainty interval easy to determine	No	No	No	Yes

4. Deciding if the data should be protected at the microdata level or the tabular level

Since tables are formed from microdata, it is possible to alter table cell values by modifying the underlying microdata. There is one fundamental advantage to modifying at the microdata level; once it is completed, any number of tables can be generated in the usual way from microdata and all such tables benefit from the disclosure protection provided by the microdata modification. Thus in situations in which a very large number of tables are to be generated and disclosure protection needs to be considered, protection at the microdata level makes sense. In addition, the protection among the full set of tables is consistent. This argument applies mainly to cases in which the microdata records after protection look like regular microdata records. When the records undergo local suppression (i.e., when some variable values are blanked) or when uncertainty intervals are associated with the modified value, creating tables from the microdata may be complicated and microdata modification may no longer have the same advantage. The main disadvantages to microdata modification are (i) the resulting table(s) probably will not be additive and (ii) the process affords little control over which cells will be modified. Thus certain non-sensitive cells, e.g., marginals, may be modified, even when there is an interest in releasing the original values. Another, perhaps less important problem is that some of the resulting cells, might, by chance, may be perturbed too little, or, on the other hand, be perturbed too much. This last problem can usually be avoided with judicious selection of the perturbation mechanism.

If only a single table is undergoing disclosure processing, the main advantage to modifying at the microdata level does not apply. It may well turn out that modifying at the tabular level will be easier. Cell suppression basically involves protection at the tabular level; however in the case of economic data, a small amount of information from the microdata may be required. This situation arises if one is trying to protect the table at the “company level.”

“Protecting at the company level” is often the goal of disclosure control with business data. Let us explain this idea. Assume each company that responds to a survey has one or more establishments. Of course, the company knows the data for all of its establishments. The establishment data contributions often appear in many different cells. The goal of protecting at the company level is to ensure that no table user (other than the given company) can derive a precise estimate of the given company’s total value (of the response variable). This type of protection implies that the values for the establishments will also be protected; i.e. no table user will be able to get a precise estimate of them. However, the converse does not hold; i.e. protecting all the establishment values will not ensure protection of the company value. Since we want both type of protection, the SO needs to protect at the company level. This requires use of some microdata since cell values are typically the sum of contributions; each contribution represents an establishment and each microdata record contains its establishment’s company identifier.

5. Deciding how to protect tabular data at the microdata level

Another term that is sometimes used for such protection is “source data perturbation” (ref: WdW2, p.36, CW). Methods of this type include: deterministic methods; topcoding, local suppression and noise addition methods (ref: EZS).

For count tables, disclosure protection involves trying to reduce the chances of record identification or attribute disclosure as illustrated by the examples in section 1. For such tables, a method such as swapping make sense if it involves splitting the link between sensitive variables (e.g., INCOME) and identifying variables (e.g., AGE and RACE).

For magnitude tables, a method that involves perturbation of the magnitude variable(s) makes sense. In this situation, the goal is not to reduce the identification risk or attribute disclosure risk but simply to ensure that the uncertainty interval associated with the magnitude variable is large enough to satisfy

confidentiality requirements. A good example of such a method is a noise addition method developed a few years ago (ref: EZS, WdW2, p.240).

6. Deciding how to protect tabular data at the tabular level

Count Tables

Recall that with count tables the disclosure issues involved identification of individual respondents (e.g., persons) or attribute disclosure for an individual or a group of individuals.

For the case of census data, such disclosures usually arise when cell counts are 1 or 2; they could arise for somewhat larger values as well. Thus an SO should use an SDC method that hides knowledge of such small cell values. One method that has long been used is rounding to a base slightly higher than the largest problematic cell value (e.g., rounding to base 3 or to base 5). Conventional rounding, in which each cell is rounded to the nearest multiple of a fixed base, independently of other cells, is very easy to implement but often produces a table that is not additive and may not provide adequate protection even if it is (ref: WdW2, p.222). To correct for lack of additivity, one could recompute marginals but then it is likely that some of the marginals will not be close to either adjacent rounded value. Thus, a challenging problem arises: is it possible to round each cell value to one the adjacent multiples of the base and still preserve additivity. The method of controlled rounding provides the solution, at least for 2D tables. There are several ways to implement controlled rounding; e.g. using either network flow, mixed integer programming, or simulated annealing. See (WdW2, p. 227) for a fairly detailed description of each approach. There may be situations in which additivity of the table is not important. Then the method of random rounding provides adequate protection and can be easily implemented.

Magnitude Tables

Rounding.

For certain types of magnitude tables, rounding may be applicable. Since rounding perturbs cell values by amounts that are the same order of magnitude, it is really not effective if the initial cell values are of different orders of magnitude. Unfortunately, it is often the case that economic activity variables, (e.g., costs, sales, profits, etc.) do vary greatly from respondent to respondent in the microdata and from cell to cell in tables derived from the microdata. However, in situation in which the response variable is of the same order of magnitude (e.g., number of employees in fast food restaurants) rounding may be a sensible method to use.

Cell Suppression.

Cell suppression has been thoroughly discussed in a number of SDC sources (ref: J,WdW2, M1). It has been implemented in various ways; most commonly using techniques from the field of operations research. These include a very fast method called network flow and a somewhat slower, but more general and better known method called linear programming. Some of the currently used algorithms for cell suppression often

provide the uncertainty about sensitive cell values that we wish to create, but occasionally they do not. Therefore, separate programs, called suppression audit programs are sometimes run to check on the performance of the suppression program. These issues have been discussed in the cited references.

There is a major drawback of using a suppressed table from the table user's point of view. This drawback is that the user, initially, seems to receive no information about value of a suppressed cell even though it is perfectly acceptable for the SO to release a rough estimate of the value. Thus, it seems like the SO has destroyed too much information. By using some simple algebra and exploiting the additivity of the table, the user may be able to derive some crude estimates for the cell values. His estimates may be so rough that they are of no use to him. However, if the user is able to access an audit program (or has the time and resources to write his own), he may be able to recover more of this seemingly lost (i.e., suppressed) information. However, cell suppression often creates a wider uncertainty interval than the required protection interval and the estimate from the audit program will reflect that. This excessive loss of information is what has motivated the development of other SDC methods in recent years.

Controlled tabular adjustment (CTA).

This method was developed in 2002 by Larry Cox and Ramesh Dandekar (ref: CD). The goal of this method is to release a reasonable value for each cell, to perturb (i.e. change slightly) the value of each sensitive cell, either up or down, by the desired protection (at most $p\%$ of the initial cell value where p is usually at most 20), and to perturb just enough of the non-sensitive cells to produce an additive table. The perturbation of the sensitive cells by $p\%$ is very similar to what is done in cell suppression, but the big difference with CTA is that the perturbed value itself is published. In cell suppression the perturbation (i.e., the change in cell value) is used inside the program to generate uncertainty cycles and the associated suppression pattern, but it is not used to adjust released values. Another big difference is that the additional non-sensitive cells, which are analogous to complementary (or secondary) suppressions, are allowed to change only by a small amount (roughly $p\%$ of its cell value). The perturbed values for the non-sensitive cells are also published in place of their initial values. If one identifies the perturbed cells in some way, one sees a perturbation pattern that is analogous to the suppression pattern one sees in a table after cell suppression. The perturbation pattern would, in general, involve more cells than the suppression pattern for a given table and value of p . Whether the perturbed cells would be identified in some way to the table user (e.g., by highlighting, by inserting a special symbol next to the cell value, etc.) probably could be decided by each SO. The SO must also decide how much information to provide the table user about the maximum allowed for a perturbation of each cell. If this information is provided, the user could easily calculate uncertainty intervals for each cell that is identified as perturbed.

Variable Base Rounding.

Gordon Sande (ref: S) has distributed a draft of a paper that described a method that he calls “variable base” rounding or “high base” rounding. He views this method as a variation of cell suppression but the table that is released has no suppressions; instead it has a rounded value for some of the cells. The rounding is designed to (1) incorporate the right amount of uncertainty (2) to make it apparent to the user that the cell is being protected. The majority of the Sande’s paper involves a careful analysis of the protection needs of typical business survey data and a comparison of the advantages of this method over earlier methods for such data.

Providing Uncertainty Intervals.

The author of this paper is currently exploring another method which is similar to variable base rounding in many ways but which publishes a value not as a single rounded value but as a value with uncertainty that is expressed in the form that is common in discussions of statistical error: value +/- error. Thus for cell ‘i’, the SO would publish perturbed values in the form {value(i) +/- uncertainty(i)} since the uncertainty depends on ‘i’. The uncertainty will be a function of the set of flows that go through a given cell (e.g., the maximum flow).

Adding Noise.

The last three methods discussed involve perturbing the cell value but they do so using deterministic (i.e., non-random) methods. The noise approach described here, in contrast, involves, using a random number generator to simulate a specified probability distribution which is sampled and the resulting value is either added to the value or multiplies the values of each of the cells to which we wish to add noise. One approach is to add noise only to interior cells and then recompute marginals to ensure that the final table is additive. Unfortunately, this will sometimes change the marginals significantly. Another approach, which is discussed in WdW (2001; p. 220) is to add the noise only to the marginals and then to distribute the noise among the interior cells using iterative proportional fitting. Thus additivity is ensured but it’s possible that some interior cells may not get adequate protection.

Deciding if an audit program is needed.

For certain methods, it has been proven mathematically that the desired protection for each cell can be achieved. However, even in this situation, using an audit program may be useful to reveal other problems, such as symmetric uncertainty intervals (possibly leading to vulnerability to a midpoint attack). For example, cell suppression based on network flow methods or more general linear programming methods will always create uncertainty (i.e., protection) intervals of at least the desired width for simple 2D tables. This can be shown mathematically. However, sometimes an uncertainty creating method is used for which there is no guarantee that the desired uncertainty will be created. Such methods are used if they are very fast and do, in practice, create the desired uncertainty a high percentage of the time. An example of this is network flow based suppression when used for 3d or higher dimensional tables. For such programs, use of an audit program should be considered.

7. Linked tables

One of the most difficult problems in tabular SDC is the protection of linked tables. The simplest way in which two tables can be linked is if they have one or more cells in common. A more general definition is: two tables are linked if there is a specific relationship between certain cells in one table and cells in the other. Such linkage can lead to disclosure problems. For example, one table may give sales for shoe stores located in cities proper and another table in metropolitan areas of those same cities. If one knows that there is only one store that is in the metropolitan area but not in the city proper, one can derive the sales figure for that store by subtraction. Tables from different surveys may be linked and tables released by different SO’s may be linked. Tables retrieved from an online database may be linked and for such a database, the user may be able to generate a very large set of linked tables that together have a high disclosure potential. Tables from a longitudinal survey are by definition linked. Linkages sometimes exist not between just two tables, but among a set of tables; e.g., an additive relationship such as

$$\text{Table A} + \text{Table B} = \text{Table C.}$$

Most of the SDC methods described in previous sections would require a separate analysis and a new implementation before they could be applied to a given type of linked tables. However, certain existing implementations are designed to handle linked tables providing all the tables are processed on a single run. In that case, the method of backtracking can be used to ensure that cells requiring protection receive adequate protection in each of the tables in which they appear. This method has some drawbacks (e.g., time requirements, lack of a guarantee of success, restriction to simple linkages, etc.) but it appears to be acceptable for the time-being as implemented in certain cell suppression programs. As currently defined, CTA, in which all sensitive cells are protected simultaneously, probably will be difficult to extend to the case of linked tables.

8. A Proposed Decision Making Process

Decision 1: Do the tables being processed need to undergo disclosure processing ?

This may depend on the sampling design (e.g., the sampling fractions in the strata).

For count data, do the tables have sensitive variables ? Could they be linked to other tables with sensitive variables ? For magnitude tables, are there response variables that need protection ?

Decision 2: What data require protection ? What type of uncertainty should be created ?

Examples: For count tables, sufficient uncertainty should be created to prevent identifications or attribute disclosures. For magnitude tables, sufficient uncertainty

should be created to ensure competitors of a company cannot estimate that company's sales figures very accurately; e.g., requiring at least p% uncertainty. For business data, need to decide whether protection should be at the company level or just at the establishment level.

How much uncertainty do you wish to create? (e.g., for count data; what should the count threshold be?) (e.g., for magnitude data; what percent (p%) uncertainty should be chosen?)

Are the tables linked? If so, in what way?

Do subject matter specialists prefer that certain cells not be modified by the SDC method (if possible)? (e.g., should the marginal cells be fixed?)

Decision 3: Does data modification at the microdata level or at the tabular level makes more sense? This depends in part on the number of tables and linkages among them.

Decision 4: How do the likely table users plan to use the tables?

For example, will there be simple uses such as lookup of cell values; or will there be statistical models built from the tables? If the latter, what types of models? If there are a number of uses, try to determine what methods are of greatest interest among the user community.

Decision 5: Which disclosure methods are the strongest candidates?

Deciding among two or more appropriate methods may require reading some papers that provide an overview and/or an analysis of various SDC methods from a table user's point of view. Papers which do a comparison of methods are especially helpful (ref: S, SG, RKG).

Decision 6: Which implementation of the method should be used?

The answer may depend on the table size as well as the data type. For magnitude tables, here are some approaches that have proven very useful. For cell suppression expressed modeled as an Integer Programming (IP) optimization problem: (1) Solve the IP problem exactly using general IP software. This will produce the true optimal answer but such an approach may not be fast enough to meet the SO's constraints. However, if one uses a customized IP solver, i.e., one that incorporates the structure of the model, significant decreases in runtimes are possible. (2) Linear programming as a heuristic; a relaxation of the IP problem and often a good approximation. For

Controlled Tabular Adjustment modeled as an IP optimization problem: (3) IP problem solved with use of a meta-heuristic (e.g., may use simulated annealing; tabu search, etc.) (4) IP problem solved using an LP heuristic.

Decision 7: Which specific software should be used?

Should the office write its own software? Or should it use existing software? Certain statistical offices have software which is distributed free of charge. It may be downloadable from a website or available on a CD. Example: see <http://neon.vb.cbs.nl/casc/> for a discussion of a downloadable package Tau-Argus (version of Argus for tables). Example: see <http://www.fcs.gov/committees/cdac/cdac.html> for links to U.S. agency software in near future.

In addition, there are private consultants who have software or services for sale.

9. Conclusion

We have presented an overview of the decision making process for selecting suitable tabular SDC methods for protecting a given set of tables. We have discussed specific methods with the goal of illustrating the type of analysis that the SO needs to undertake. Such analysis considers the general characteristics of a method before concentrating on more detailed implementation issues.

To make the best possible decision about which method to use, the SO really needs quite a bit of information, some of which may be difficult to acquire. In addition to this specific knowledge, there may be some general research which, in the future, would help the SO in the comparison of SDC methods. We mention three topics which have a strong applied flavor. They involve the way users use the tables, errors in the data, and table linkages, respectively.

The broadest question discussed in this paper is determining the impact on table users of a decision by a statistical office to use a given SDC method to protect a given set of tables. As mentioned above, one difficult aspect is that there are likely to be a wide range of uses of the tables being protected, ranging from a simple lookup of a few cell values to development of complex statistical models. Thus the first step of this analysis involves collecting data from the users (or people who interact with them) on how they use the tables. We suspect that many challenging statistical problems would arise in this analysis; some could be quite interesting. For example, one might explore the effect of CTA perturbation on tables of counts when the user is constructing a certain class of log-linear models.

It may be possible to relate the uncertainty in cell values, that a given SDC method creates, to uncertainty in the coefficients of a log-linear model. If there are a small number of such models and other table uses to explore, this task would be manageable.

Another important topic is the role that data error analysis plays in the use of an SDC method. This topic, which depends heavily on traditional survey methods, would allow the office to compute the effect of sampling and non-sampling errors on the setting of protection levels. For some methods, it may be easy to adjust the amount of protection to reflect the existing data error. Of course, this assumes that the office is able to compute at least rough estimates of the data errors. Such estimation is often difficult.

The study of how linked tables are to be protected is one that needs to be addressed for each new SDC method. Certain methods can easily extend their protection from a single table to linked tables; at least if all the tables are processed on a single run. However, other methods, either cannot handle linked tables at all or at least not easily. Even in cases when the theoretical description of the algorithm indicates that the method handles linked tables, the current software for the method may not have that capability. For those methods that can handle a given set of linked tables on a single processing run, there remains the issue of how to handle linked tables that are processed on different runs. Perhaps some general results about handling linked tables could be developed so that each new SDC method could be evaluated quickly with respect to this capability.

As is often the case when comparing statistical methods, one must have good knowledge of the data at hand and the uses to which the statistical products (e.g., tables) will be put, before deciding which method is best to use. When a specific method has been selected or is at least a strong candidate, there are practical considerations such as the mathematical or statistical approach used in the implementation and the specific software to use. Thus there is much information to gather and analysis to be done, if the SO wishes to select the best tabular SDC method to use with a given set of tables.

Acknowledgments. I would like to thank my colleagues at the Census Bureau, especially Laura Zayatz, and Jim Fagan, as well as Jose Dula of Virginia Commonwealth University, and Steve Roehrig of Carnegie Mellon University for discussion of various topics in this paper.

References

CD: Cox, L.H., and R.A. Dandekar (2002), "A Disclosure Limitation Method For Tabular Data That Preserves Data Accuracy and Ease-of-use", presented at Federal Comm. Stat. Methodology Conf., Nov. 2002.

CW: Cuppen, M., and Willenborg, L.(2003), "Source Data Perturbation and consistent sets of safe tables", *Statistics and Computing* v.13, pps. 355-362.

EZS : Evans, T., Zayatz, L., Slanta, J. (1998), "Using Noise for Disclosure Limitation Establishment Tabular Data", *Journal of Official Statistics*, December, 1998.

J: Jewett, R.(1993), "Disclosure Analysis for the 1992 Economic Census", unpublished Census report 1993.

M1: Massell, P. (2001), "Cell Suppression and Audit Programs used for Economic Magnitude Data," Statistical Research Division report, U.S. Census Bureau, <http://www.census.gov/srd/papers/pdf/rr2001-01.pdf>

M2: Massell, P. (2004), "Statistical Disclosure Control for Tables: Determining Which Method to Use", *Proceedings of Statistics Canada's 20th International Methodology Symposium* held October 2003, (to be published).

RKG: Russell, J. N., J. P. Kelly, and F. Glover (2002), "The Bureau of Transportation Statistics' Statistical Disclosure Limitation Method for Tabular Data: A Review", 2002 Proc. Amer. Stat. Assn., Gov. Sect. [CD-ROM], Alexandria, VA.

S: Sande, G. (2003), "A Less Intrusive Variant on Cell Suppression to Protect the Confidentiality of Business Statistics", unpublished report.

SG : Salazar-Gonzalez, J.J. (2004), "A Unified Mathematical Programming Framework for different Statistical Disclosure Limitation Methods for Tabular Data", to be published in *Operations Research (J. of INFORMS)*.

WdW1: Willenborg, L., and T. de Waal, T.(1996), *Statistical Disclosure Control in Practice*, Lecture Notes in Statistics, v. 111, Springer, 1996.

WdW2: Willenborg, L., and T. de Waal (2001), *Elements of Statistical Disclosure Control*, Lecture Notes in Statistics, v. 155, Springer, 2001.