Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2003: Challenges in Survey Taking for the Next Decade

2003

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2003
Challenges in Survey Taking for the Next Decade

STATISTICAL DISCLOSURE CONTROL FOR TABLES: DETERMINING WHICH METHOD TO USE

Paul B. Massell[1]

ABSTRACT

The development of some new statistical disclosure control (SDC) methods for tables in recent years provides statistical agencies with additional ways to protect the confidentiality of their data. However, it also means that these agencies need to make choices about which method should be applied to a given table or set of (possibly linked) tables. In this paper, we analyze several important SDC methods with respect to a set of important factors that an agency needs to consider in choosing a method. The dominant factor in this decision process should be the nature of the confidentiality pledge that an agency is required to enforce. Another important factor is the development time and runtime of the software associated with each method. Another factor, one which has received too little attention, is the way in which the tables will be used by table users. Of course, this factor is difficult to analyze since different users use the tables in different ways; e.g., some users simply want to look up a few cell values; others are interested in sophisticated statistical modeling. In this paper we propose some statistical methods for analyzing SDC methods with respect to these factors. Hopefully, this analysis will help agencies select the best SDC method to use with a given set of tables.

KEYWORDS: Cell Perturbation; Cell Suppression; Confidentiality; Statistical Disclosure Control; Statistical Tables.

1. INTRODUCTION[2]

Most survey and census data are collected with a pledge of confidentiality. In this paper we assume that the statistical office has translated that pledge to a fairly precise set of confidentiality rules for determining whether a table or set of tables contains confidential information. Suppose then that the statistical office has determined, possibly through the use of software, that a set of tables may contain confidential information and should therefore undergo some type of disclosure processing. There are then several factors the office should consider when deciding which statistical disclosure control (SDC) method(s) to apply to the set of tables before releasing them to the public. Selection of a method involves several steps. Firstly, the office must have knowledge of the range of tabular SDC methods that exist. Many of these have been described in statistical journals, technical reports, or books (Willenborg et al, 1996, 2001). These sources often describe the mathematical and statistical properties of a tabular SDC method, the history of its development, development of software which implements it, and the results of computational experiments. The computational experiments are important for revealing limitations of the method and for giving indications of the processing times required for tables of various sizes. Both the statistical properties of the method and the properties of the software will determine whether the given SDC method is suitable for protecting a given set of tables.

We begin by defining the "forward process" for a tabular SDC method in section 2 and the "backward process" in section 3. Section 4 deals with the type of data that appears in the table and the candidate SDC methods for that type of data. In section 5 we compare various types of uncertainty that can be created to prevent disclosure. Section 6 deals with the usefulness of tables after disclosure processing. Section 7 proposes a decision making process for dealing with all the above issues.

[1]Statistical Research Division, U.S. Census Bureau, 4700 Silverhill Rd., Rm. 3209-4, Washington, D.C. 20233, U.S.A., paul.b.massell@census.gov

[2]This 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(s) and not necessarily those of the U.S. Census Bureau.

2. DEFINING THE FORWARD PROCESS FOR A TABULAR SDC METHOD

We have found it useful to define the terms Aforward process@ and Abackward process@ for an SDC method. The **forward** process for a tabular SDC method is what is normally considered to be the key step of disclosure processing; e.g., a statistical office runs a suppression program or a perturbation program on the table(s). The goal is to produce a set of tables that have been protected with respect to disclosure. In general, this involves altering some cell values. It consists of several substeps.

A. Using a sensitivity rule such as the p% rule or the (n,k) rule (Willenborg, 2001), identify those cells that are sensitive. Denote this set of cells by S-cells. In the case of linked tables, determining a cell=s sensitivity may be complicated. (If S-cells is empty, no SDC method need be applied.).

B. For each sensitive cell, define the minimum amount of uncertainty (MinU) that the SDC should create. MinU may be viewed as a function on S-cells to some range set MinU(S-cells) that expresses uncertainty. One common way of expressing uncertainty is an interval. In this case MinU(cell i) would be an uncertainty interval for cell i. Ideally, it would be useful to define the maximum amount of uncertainty that the SDC should create for each cell (not just S-cells). MaxU may be viewed as a function on (Cells) with range set MaxU(Cells).

C. Apply the SDC method to the table with S-cell, while satisfying the constraints imposed by MinU(S-cells) and MaxU(Cells). The SDC processing of a table is deemed to be successful when a new table has been created satisfying these uncertainty constraints. Determining if the processing was successful may require use of a backward process (see below).

D. If the method is stochastic in nature, it may be possible to express the procedure in terms of a set of probability functions (pf-s) (or densities); one associated with each cell. They have the form: Pr(v-post = y | v-pre = x) where v-pre is the cell value prior to (disclosure) processing and v-post is the cell value after processing. We call these pf-s (or densities) **forward** probabilities. The simplest examples of these are in stochastic rounding (Willenborg, 2001, p.224) or in random perturbation of count tables. In these cases, MinU(S-cells) and MaxU(Cells) are pf-s rather than intervals.

3. DEFINING THE BACKWARD PROCESS OF A TABULAR SDC METHOD

The backward process is performed on the table after the forward process for some SDC method that has been performed. An example of a backward process is a disclosure audit program which is run after cell suppression has been done. If backward processing is performed by the statistical office, it is generally done prior to release of the table. In this case, its purpose is to determine if the forward processing was successful and perhaps to note some detailed information about the processing such as specific levels of uncertainties for certain cells (possibly all). If the forward processing is successful, the table is released to the public. After the table is released, backward processing may be applied by any data user (with some computational resources). His/her purpose may be to reconstruct the original (unaltered) table as closely as possible, perhaps in preparation for some modeling effort. For example, if a table is released with some cells suppressed, a modeler may need to impute values for these cells prior to modeling (see below). Of course, the purpose may be more limited, simply to calculate the best estimate, that the table allows, of a few cell values. Calculating the best estimate involves calculating the uncertainty interval or the pf (or another uncertainty descriptor) associated with a cell value and then creating a single value, e.g., the mode of the density, from that information.

Observations about backward processing

A. Some tabular SDC methods are guaranteed to be successful; i.e., to meet all the uncertainty requirements. This guarantee is usually based on a mathematical result. For example, cell suppression programs based on a network flow algorithm and run on simple 2D tables, will provide (at least) the amount of protection that they purport to provide. For these, backward processing is not needed by the agency to check Min-U(S-cells) although it might be useful to assess Max-U(Cells); that is to see if too much uncertainty is being created. For the cell suppression example above, theory guarantees that the method will not undersuppress but does not rule out oversuppression.

B. If the forward processing creates an uncertainty interval for each cell, backward processing can check that the interval is suitably located and wide enough to meet the desired uncertainty.

C. If the forward processing is stochastic in nature, and the forward probabilities can be clearly specified, one may be able to use Bayes= formula to compute backward probabilities. Backward probabilities can be used for assessing the amount of the uncertainty created by a SDC method.

Backward probabilities have the form: Pr(v-pre = x | v-post = y) where v-pre is the cell value prior to (disclosure) processing and v-post is the cell value after processing. These probabilities should have an uncertainty large enough to satisfy MinU(S-cells).

4. TYPE OF DATA AND CANDIDATE TABULAR SDC METHODS

The first step in selecting a tabular SDC method is to identify the type of data in the given table or set of tables.

A. What type of data are in the table? Magnitude data or frequency count data?
B. If the data are from a sample survey, can the sampling error be estimated for each cell?
C. How many tables need to be constructed? What are their sizes? Are they hierarchical?
D. Are there cells that appear in two or more tables? If so, the tables are linked. Are they linked in some less obvious way; e.g., via an additive relationship among the tables?

We now discuss how the answers to these questions affect the choice of SDC method.

A1: (Frequency) Count Data

If the tables contain count data, the key disclosure problem is the presence of cells with very small counts. Some agencies consider cells with counts below three to be problematic; in that case an agency must use some SDC method that eliminates those low counts. Table redesign involves redefining categories for the row and/or column variables. Collapsing of categories along any dimension is a simple example of table redesign in which one combines two or more categories into one. Rounding also can be used for count data. Rounding changes almost all the cells of a table so, in a way, it is Aoverkill@. In addition, rounding has some implementation challenges; e.g., conventional rounding does not usually preserve additivity; other types of rounding preserve additivity by allowing non-nearest rounding for the marginals (i.e., the marginal is not necessarily equal to either of the two nearest multiples of the base). Controlled rounding does preserve additivity but it is time-consuming and is not guaranteed to work in dimensions three or higher. Random perturbation is also an option; it is quick but also suffers from Aoverkill@. Newer methods of perturbation, such as controlled tabular adjustment (CTA) (Cox, Dandekar, 2002) are also options. One implementation of CTA (Russell et al, 2003) allows the user to declare any cell with a count less than ›n= to be sensitive. It then changes cells counts that lie in the interval [1,n-1] to either 0 or ›n=, often assigning the nearer of these two values. Additivity is preserved, in general, and there is an option to request that the program try to achieve additivity without changing marginal values. Finally, suppression can be used (Willenborg et al, 2001, p.34), but Sande (2003) says it is not normally recommended for count tables since it does not satisfy the type of protection required for counts.

A2: Magnitude Data

If the tables contain magnitude data, each cell represents the sum of a variable for the set of respondents defined by the spanning variable values for the cell (e.g., the row and column variables). For example, with economic magnitude data, the variable might be sales (in dollars) and the respondents might be (business) establishments that sell a certain product (row) and are located in a certain city (column). With economic data, the count of establishments associated with a cell is considered publicly available and thus no attempt is made to protect it. It is the response (i.e., magnitude) variable value associated with a cell that has the potential for being sensitive. The goal is to make it hard to estimate the value accurately; so that the value (e.g., sales) for any one respondent cannot be estimated accurately (e.g., within 10%) by any table user, in particular by a survey respondent whose data also contributed to the cell value.

Cell suppression can be used to create at least a specified amount of uncertainty in such cells. The uncertainty can take two forms (1) a 2-sided interval that contains an interval of the form [t - a, t + b] where t is the true value and a, b are the one sided uncertainties we wish to introduce or (2) a sliding interval of width a + b that contains the true value t. Traditionally agencies do not publish the actual uncertainty interval for each suppressed cell; however any table user with the resources to construct a linear program that models the table additivity will be able to compute these uncertainty intervals (also called feasibility or audit intervals). Certain types of perturbation create an uncertainty region that may consist of two intervals of the form [t-a-c, t-a] and [t+b, t+b+d] or it may consist of only two points t-a and t+b. Traditional rounding is less appropriate for magnitude data than for count data because the magnitude data are often highly skewed. If cell values are of different orders of magnitude, rounding will likely introduce too much uncertainty to some cells and too little to others. Finally, category collapsing may be an option, but for economic data there is a desire to have consistency of row and column definitions from year to year. Recently Sande (2003) has described a new type of rounding (which he calls Ahigh base@ or Avariable base@) that is designed for magnitude data. The user initially rounds only those cells determined to be sensitive. The rounding procedure itself is somewhat complicated; it is designed to ensure that the specified amount of uncertainty is met and that a midpoint computation (i.e., taking the midpoint of the rounding interval) will not yield a value close to the true value.

B. The Effect of Sampled Data

When table cell values, whether count or magnitude, are computed from sampled data, the risk of disclosure is usually less than it would be if the data were computed from data based on the entire population (i.e., census data). Thus the uncertainty that results from a given SDC method should take into account sampling error (and ideally non-sampling error as well) whenever these can be estimated. To date, most SDC methods do not do this. To compensate for ignoring the sampling effect in a direct way, SDC users sometimes set a protection level in the program lower than what they want for the actual uncertainty. This strategy is based on the idea that the true protection is greater then the nominal (e.g., specified) protection due to data errors of various types (including sampling). In Willenborg (2001) the effect on uncertainty for count data (p.149) and the effect for magnitude data (p.144) are discussed. Sande=s variable base rounding (Sande, 2003) allows for use of any type of sampling or non-sampling error that can be estimated.

C. Number of Tables; Sizes and Substructures

It is important to know how many tables need to undergo disclosure processing and what their sizes and substructures are (e.g., hierarchies) because these aspects provide a measure of the scope of a particular disclosure processing problem. For SDC problems of large scope, time-consuming SDC methods often need to be ruled out because their runtimes exceed the time allotted for disclosure control during production runs.

D. Linked Tables

In (Willenborg, 2001, p.150), an introduction to the challenging problem of linked tables is presented. Ideally, an agency should process linked tables simultaneously in a way that incorporates all their relationships. However, due to computational considerations, it is often not possible to do that, especially for large tables. Sometimes SDC methods can handle linked tables sequentially. This approach is likely to require a method that allows for revisiting a table for additional processing after the initial visit; this is called ›backtracking=. Most of the cell suppression programs used at the U.S. Census Bureau use backtracking (Jewett, 1993). This type of revisiting occurs during a given production run. Ideally an office should consider consistent handling of linked tables across different production runs.

5. COMPARING UNCERTAINTIES; DETERMINING IF THE UNCERTAINTY IS SUFFICIENT

For some U.S. federal agencies, there are laws which require the agency to protect the confidentiality of all respondents to its surveys and censuses. General legal statements are not specific enough to lead to quantitative rules for implementing these requirements. There must be policy developed which is based to some degree on quantitative experience and/or thought experiments and/or computational experiments. Such policy ultimately needs to consider the questions: (1) for count data: what amount of uncertainty regarding the ability to identify a respondent (or to do

inferential disclosure) using the table is sufficient to say that the confidentiality of all respondents has been protected ? (2) for magnitude data: what amount of uncertainty needs to be introduced in a table so that the best estimate based on the table does not allow a table user to violate the confidentiality of a respondent ? Thus confidentiality policy converts a qualitative statement about confidentiality into a set of quantitative statements that can be used for evaluating specific statistical disclosure control. One might use the term ›quantitative confidentiality requirement for tables‹ to describe this notion.

5.1 The Forward Process: Uncertainty creation

5.1.1 Interval uncertainty: Feasibility intervals computed by a disclosure audit program

With cell suppression, we get an uncertainty interval for each suppressed cell. How should we go about calculating the uncertainty interval? Often the calculation assumes that the table is independent of other tables. However, if the table is linked, the audit program should, ideally, take that linkage into account. A difficult policy question is: how much uncertainty should be introduced to meet confidentiality requirements, e.g., is 5% sufficient (i.e., the uncertainty interval for a cell value contains the interval (0.95* v, 1.05*v))? If not, what about 10% or 20%? Should the upper and lower protections be made equal or should they be made quite different in order to make the midpoint guess strategy less successful? Of course, even when zu and zl are made equal (for the required uncertainty), the actual uncertainty created by a program can be quite asymmetric about the true value. Another approach is to implement sliding protection. This protects against the midpoint guess strategy and since it is a less constraining protection requirement, there are often fewer suppressions (Kelly et al, 1992).

5.1.2 Interval Uncertainty: Rounding intervals

Recently, Sande (2003) discussed a way of forming protection intervals for sensitive cells that can be expressed as a type of rounding (called Avariable base@). The rounding interval has advantages over suppression. Specifically, the table user is immediately able to associate an uncertainty (protection) interval to each sensitive cell. Other cells, which have error (sampling or non-sampling) that can be estimated, can also have an uncertainty interval associated with them. Thus the table user does not have to compute any uncertainty intervals, he can simply read them along with the value. This way of expressing uncertainty is close to the traditional way of expressing an estimate in the format: x " error. The subtle aspect of the method is how to compute uncertainty protection intervals that are not symmetric about the true value. Such an approach is needed so that a midpoint strategy will only rarely yield a value close to the true one.

5.1.3 Finite uncertainty sets

With certain types of perturbation or rounding, the sophisticated table user can perform a backward analysis (see above), and derive a probability function on the set of possible values (uncertainty sets). These uncertainty sets are often finite. Now the policy question is: what does the shape of the probability function have to be to ensure adequate confidentiality protection? For example, suppose for perturbation of magnitude data, a user is able to determine that the true value is either t1 or t2; each with probability 1/2. Is that uncertainty sufficient? What about for a more skewed probability function, e.g., Pr (t1) = 0.9, Pr (t2) = 0.1? The notion of entropy as used in books on information theory might be a useful way to measure the uncertainty. [Recall this type of entropy is computed by summing over all probabilities i, of (p (i) * (-log (p (i))) where log represents log to base 2. This quantity is non-negative and is bounded above by log (n) where n is the number of non-zero probabilities associated with the cell value.]

5.1.4 Format for modified cell values

The last stage of the forward process is changing the cell values to reflect the uncertainty created. For suppression, a single letter, e.g., a ›D‹, may be placed in each cell. In traditional rounding, one simply publishes the cell value rounded to the nearest base. In more complicated types of rounding it may be suffice to publish a single value in a cell but it will likely be necessary to include Adecoding information@ about how to construct the uncertainty interval from the value.

5.2 The Backward Process: The Estimation of the Cell Value and Uncertainty

Suppose that the user is presented with an interval, either explicitly as with rounding or implicitly as with suppression. One can assume some probability density f (v) for this interval. Then one can compute any estimate of the cell value from f (v). As is common in Bayesian analysis, if one has no prior information about f (v), one might assume it is the uniform density on the uncertainty interval [a, b]. Recall in that case we have: f (v) = 1/(b-a); mean=(a+b)/2; var=$(b-a)^2$ /12.

5.3 Example of the Forward and Backward Analysis of an Algorithm: Case of Controlled Tabular Adjustment (CTA)

5.3.1 Uncertainty Creation

Step 1: For each sensitive cell (with value v), one chooses either upward or downward change, each with probability equal to 1/2; i.e., either v1 = (1+p/100)*v or v2=(1-p/100)*v.
Step 2: Use an optimization model similar to that used for suppression, similar in the sense that it solves for cell changes in an additive table and has bounds for the perturbations.

Unlike many commonly used suppression programs which protect sensitives sequentially, in CTA, all sensitives are set at their protection values simultaneously. Thus only one flow need be computed (e.g., only one call to the linear programming or integer programming solver). Nonzero flows through non-sensitives are added to current values. Bounds on the flow are usually set to be a small percent of cell value (so perturbations will be small).

5.3.2 Changing of Cell Values

For each sensitive cell, change to either v1 or v2 with probability of 1/2. Each non-sensitive cell is changed to a value in its bounds interval. In general, the bounds for perturbations need to be greater than the changes required for sensitive cells. The choice of the bounds multiplier is dependent on what size for the perturbations render the cell value useless. Also, a higher bound for non-sensitives cells means that fewer of them need be perturbed to protect the sensitive ones.

5.3.3 Uncertainty Estimate

If the user knows the value of p, and he knows the cell was sensitive, then he needs to guess whether the cell was perturbed up or down. Thus if vn=v1 (perturbed up) then va=vn/(1+p/100) or if vn=v2 (perturbed down) then va = vn/(1-p/100) where va=estimate using vn and the (up or down) guess. Uncertainty is actually larger than this since (1) the user may not know ›p‹ or (2) he may not know if the cell is sensitive, or was modified at all.

6. DETERMINING THE USEFULNESS OF THE TABLES AFTER DISCLOSURE PROCESSING

Since an agency›s purpose in releasing tables is to allow the tables to be used for various purposes by (a possibly large number of) users, the effect of disclosure processing on various sets of tables should be explored. There are various statistical experiments that could be performed.

In the analysis below, we use the word ›model‹ in a general sense; it includes simple uses of tables (e.g., reading the values in designated cells to make simple comparisons or calculations) as well as traditional statistical models (e.g., log-linear models).

6.1 Statistical experiments to determine the effect of SDC on a model built from a table

Statistical Experiment 1: Determine the effect of a given SDC method on a given type of model

Suppose an agency currently uses only one SDC method, denoted D1, on, say, magnitude tables. Suppose a table user computes a certain type of model, denoted M, based on such a publicly released table; i.e., one that has undergone disclosure processing. Denote the specific model so developed by M(Post-D1). Suppose the table user acquires special status that allows him to visit the agency and do the same type of modeling on the table as it exists just prior to disclosure processing. Using this pre-disclosure table, he produces a model M(Pre-D1). Denote by Adist@ some sensible way of measuring the distance (or difference) between these two specific models of type M. For example, Adist@ might measure the change in the parameter values. Then the distance (or difference) between the two specific models just described is denoted by:

dist(M(Pre-D1), M(Post-D1)).

If this difference is significant, we can conclude that method D1 has a big impact on the modeling effort of type M for the given table. The modeler might feel he needs to use the Pre-D1 table or a table that is disclosure processed using another method.

Statistical Experiment 2: Compare the effect of two SDC methods on a given type of model

Suppose for a given table, a modeler wishes to construct a model of type M. Suppose he has access to two versions of a given table. The pre-disclosure table is the same for both tables. This common table was then disclosure processed using two disclosure methods D1 and D2. The user would want to use the method Di that minimizes dist(M(Pre-Di), M(Post-Di)) over i = 1,2. (As above, we assume that the user has special status that allows access to the original table.)

6.2 Table Usage and SDC methods

6.2.1 Simple Usage: Reading some cell values to make simple comparisons or calculations

6.2.1.1 If suppression was applied to the table

A user has one of two extreme cases:
(1) statistical office=s best estimate
SDC-Unc(v) = 0 where SDC-Unc(v) is the uncertainty interval for the value v.
(2) No direct value; only suppression symbol
SDC-Unc(v) = Unc(v=s suppression interval) where the r.h.s. is defined as the uncertainty associated with the suppression interval for v.

6.2.1.2 If CTA perturbation was applied to the table

We have: 0 # SDC-Unc(v) # Est(Max-Pert) where Est(Max-Pert) is the user=s estimate of the largest perturbation applied to any cell. Usually Est(Max-Pert) n Unc(v=s suppression interval). However, the user has a non-zero uncertainty for each cell, even those that have not been perturbed.

6.2.2 Modeling

6.2.2.1 If suppression was applied to the table

Many models cannot be developed with tables that have blank cells (i.e., missing data), i.e., many modeling methods require complete tables. One can impute missing (e.g. suppressed) cells in various ways using an algorithm based on maximizing likelihood (e.g. iterative proportional fitting) or maximizing entropy. However, much research is needed here to determine the effect of blank cells on the final model.

6.2.2.2 If CTA Perturbation was applied to the table

In this case, the table is complete but there is a small but unknown amount of uncertainty about each cell. The question here is how great is the effect of the perturbations on the final model. One needs to conducts statistical experiments like those described above.

7. A PROPOSED DECISION MAKING PROCESS

1st decision:

Describe the tables about to undergo disclosure processing.
What data are you trying to protect?
What type of uncertainty are you trying to create?

(e.g.,for a count table; sufficient uncertainty to prevent inferential disclosures?)
(e.g.,for a magnitude table; try to ensure that competitors of a company cannot estimate that company's sales figures very accurately; say, at least 10% uncertainty)

Do you want to protect data 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 uncertainty should be chosen?)

Are the tables linked? If so, in what way?
Are there preferences by subject matter specialists within the office that certain cells not be modified by the SDC method if possible?

(e.g., should the marginal cells be fixed?)

2nd decision:

What type of tabular SDC method could be used for the type of data that you have?
(1) table redesign; e.g., collapsing categories
(2) rounding (various types)
(3) suppression
(4) perturbation (various types)

3rd decision:

How is the table going to be used?

(e.g., 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?)

4th decision:

Taking into account the decisions above, which methods appear to be the strongest candidates?

This may require reading books or papers that provide an overview and/or an analysis of various SDC methods from a user's point of view. Papers which do a comparison of methods are especially helpful (Salazar-Gonzalez, J.J.,2002), (Russell et al, 2002), (Sande, 2003).

5th decision:

Which implementation of the method should you use? The answer may depend on the table size as well as the data type.
(1) Integer programming (IP); using a general callable routine; often yields best answer but may not be fast enough to meet time constraints
(2) Integer programming; customized to the problem at hand; can be much faster than general IP (One might call this Smart IP or Fast IP; it uses branch-and-bound or branch-and-cut methods)

(3) Linear programming as a heuristic; a relaxation of the IP problem and often a good approximation
(4) Meta-heuristic (e.g., may use simulated annealing; tabu search, etc.)

6th decision:

Which specific software to use?
Should you build your own software? Or can you 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.fcsm.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.

8. CONCLUSION

What research would facilitate a better comparison of SDC methods?

We mention three factors 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; ranging from a simple lookup of a few cells 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 specific 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 factor 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 protection levels 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 sometimes 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.

In this paper, we have included only a brief discussion of those factors that focus on the algorithmic aspects of SDC methods and their software implementations. Fortunately, these topics are extensively discussed when a new SDC method is introduced or new approaches to an existing one are discovered.

ACKNOWLEDGMENTS

I would like to thank my colleague Jim Fagan of the Census Bureau, Jose Dula of the University of Mississippi (who visited Census for a year), and Steve Roehrig of Carnegie Mellon University for discussion of many of the topics addressed in this paper.

REFERENCES

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

Cox, L.H. and J. Kelly (2003), @Balancing Data Quality and Confidentiality for Tabular Data@, Proceedings of the 2003 Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Luxembourg.

Dandekar, R.A. (2003), ACost Effective Implementation of Synthetic Tabulation (a.k.a. Controlled Tabular Adjustments) in Legacy and New Statistical Data Publication Systems@, March 2003, unpublished report.

Dandekar, R.A. and L.H. Cox (2002), AControlled tabular adjustment: an alternative to complementary cell suppression for disclosure limitation of tabular data@, unpublished report.

Dula, J., Massell, P., Fagan, J., (2003), AStatistical Disclosure Control for Tabular Data@, unpublished report.

Evans, T., Zayatz, L., Slanta, J. (1998), AUsing Noise for Disclosure Limitation Establishment Tabular Data@, Journal of Official Statistics, December, 1998.

Fienberg, S. (2003), AStatistical Disclosure Limitation: Releasing Useful Data for Statistical Analysis@, presented at Bureau of Transportation Statistics, April 28, 2003, http://www.bts.gov/confidentiality_seminar_series/2003_04/

Fischetti, M., J.J. Salazar Gonzalez (2000), AModels and Algorithms for Optimizing Cell Suppression in Tabular Data with Linear Constraints@, J. Amer. Stat. Assn., v.95, no.451, pps 916-928; Sept 2000.

Glover, F. (1990), ATabu Search: A Tutorial@, Interfaces 20:4 July-August 1990 (pp. 74-94)

Jewett, R.(1993), ADisclosure Analysis for the 1992 Economic Census@, unpublished Census report 1993.

Kelly, J.P., Golden, B.L., Assad, A.A. (1992), ACell Suppression: Disclosure Protection for Sensitive Tabular Data@, Networks, Vol. 22 (1992), p.397-417.

Massell, P. (2001), ACell 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

Russell, J. N. and J. P. Kelly (2003), ABureau of Transportation Statistics' Prototype Disclosure Limitation Software for Complex Tabular Data@, Proceedings of the 2003 Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Luxembourg.

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

Sande, G. (2003), AA Less Intrusive Variant on Cell Suppression to Protect the Confidentiality of Business Statistics@, unpublished report.

Salazar-Gonzalez, J.J. (2002), AA Unified Mathematical Programming Framework for different Statistical Disclosure Limitation Methods for Tabular Data@, CASC workshop, http://webpages.ull.es/users/casc/#Working%20papers%20and%20articles

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

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