SUPPRESSION METHODOLOGY IN STATISTICAL DISCLOSURE ANALYSIS

Lawrence H. Cox, U.S. Bureau of the Census

INTRODUCTION

There are various classes of problems within the realm of statistical disclosure analysis and to each is associated a set of disclosure avoidance techniques. This paper is concerned with one specific disclosure avoidance technique, cell suppression, and the disclosure problems to which this technique applies. This limitation of scope does not, however, extend to the techniques we describe for analysis of the network defining the tabulation cells, as these techniques admit application in a variety of settings in and out of statistical disclosure analysis. In particular, they may be employed to define a bottom-to-top tabulation system for the network.

The suppression problem is discussed and solved here deterministically and completely within the context of the publication network, according to techniques and analyses developed by the author. This deterministic analysis is prerequisite to any associated stochastic or extra-network analysis, in particular because it provides the proper context for such analyses. The emphasis of this paper will be to highlight the relevant methodological problems posed in the application of suppression techniques in disclosure avoidance. Dut to limitations of space, it will not deal with the relevant issues and problems in the design and development of an automated system to effect these analyses and the practical experience gained from the development of a disclosure analysis system for the 1977 Economic Censuses currently underway at the U.S. Bureau of the Census.

The reader may refer to [2] for a discussion of other techniques of disclosure avoidance and to [1] for further explication of the terminology.

THE SUPPRESSION PROBLEM

To protect the confidentiality of the identity or response of each respondent to a set of statistical publications, a test of sensitivity must be applied to each tabulated cell for each statistic to be published. This is accomplished according to an operant definition of sensitive cell for this statistic. In general, a cell is sensitive for a particular statistic if the value of the cell for this statistic could be employed to yield an unacceptably close upper estimate of the contribution of any one respondent to the total cell value.. An unacceptable estimate of this response would by definition breach the confidentiality of the respondent by effectively publishing its response or providing information which could lead directly to a determination of the respondent's identity. For example, when the data are categorical (qualitative), so that each respondent contributes 1 to the cell value if the respondent is a member of the cell and 0 otherwise a threshold rule defines a cell to be sensitive if the cell contains n or fewer respondents, for n a fixed (small) positive integer. In applications involving aggregate (quantitive) data, such as the U.S. Economic Censuses, a dominance rule defines a cell to be sensitive if n or fewer respondents in the cell contribute greater than k% of the total cell value, for fixed parameters n and k; n is a small positive integer and $0 < k \leq 100$. If respondent data are assumed positive, then threshold rules for quantitative data are dominance rules with k = 100.

According to suppression methodology, the values of sensitive cells are not published (i.e., are "suppressed from publication") for the statistics for which they are sensitive. As linear relationships usually exist between tabulation cells in a publication network, upper and lower estimates of the values of the suppressed sensitive cells may be obtained by linear techniques and, in some instances, precise determination of the value of a sensitive cell may be made. As a result, a consistent definition of what constitutes an acceptable estimate of the value of a suppressed sensitive cell must be made in order that additional, appropriately chosen, linearly related non-sensitive cells, called complementary suppressions, may also be suppressed from publication. These complementary suppressions are made to insure that only acceptable estimates of the values of sensitive cells may be obtained from the network. Equally important, the complementary suppression process must be performed so as to minimize its adverse impact on the information content of the publications.

Each of the above concepts must be made precise to the extent that they may be measured in a predetermined and meaningful sense. These several issues will be dealt with in separate sections of this paper. Interrelationships between them will be discussed at appropriate points.

DEFINING ACCEPTABLE ESTIMATES OF SUPPRESSED SENSITIVE CELLS

Assuming the respondent date are non-negative, if the value of a cell or union of cells containing a particular individual respondent to a cell is known, then this value is an upper bound of the value of this respondent's datum. Similarily, zero is a lower bound on this value. In general, therefore, an interval estimate of the value of each individual response to each cell exists. Sensitivity rules are developed to identify those estimates of individual respondent data which are unacceptable according to established criteria. Acceptable estimates of sensitive cells therefore must be defined so that the estimates of the value of individual respondent data they provide conform to the corresponding estimates obtainable for respondent data from non-sensitive cells. Acceptable estimates must be determinable from the sensitivity rule and, ideally, one should be able to pass from formulae for acceptable upper and lower estimates of sensitive cells to a formula which describes the sensitivity rule.

If cell sensitivity for categorical data is defined by a threshold rule, then it follows that an unacceptable lower estimate of the value of a suppressed sensitive cell should be defined as zero, and an unacceptable upper estimate of its value should be defined to be greater than the parameter n. This results from the fact that a threshold rule is applied to categorical data to prevent any individual from being classified in a group of fewer than $n+1$ respondents.

To determine acceptable estimates of suppressed sensitive cells in a publication network of quantitative data, one must examine the available methods of estimation of cell values from above and below and the corresponding estimates of individual respondent data which can be made for respondents in non-sensitive cells. In general, dominance criteria are employed because, if there is dominance of a cell X by a small number n respondents, then it is possible for one of the dominating respondents to subtract its contribution from the total cell value $V(X)$, thereby obtaining an undesirably close upper estimate of the total value of the responses of the other dominating respondents, and thereby a refined upper estimate of the contribution of each of these other (n-1) dominating respondents. Indeed, it is the value of $D(X)$, the total contribution of the n largest respondents, which must in general be protected. If X is sensitive, $V(X)$ is suppressed only because it represents an unacceptably close upper estimate of $D(X)$. For cells X in which the total contribution $D(X)$ of the n largest contributing respondents lies below the dominance threshold (i.e., $D(X) < (k/100)V(X)$), $V(X)$ is by definition an acceptable upper estimate of the value of the response of any of the n dominating respondents. In particular, this is true when $D(X) = (k/100)V(X)$, in which case publishing $V(X)$ protects $D(X)$ by $((V(X)-D(X))/D(X))\%$ of its value, i.e., by $((100-k)/k)\%$ of the value of $D(X)$. For sensitive cells, therefore, it is reasonable to define an acceptable upper estimate of the value $V(X)$ of a sensitive cell X to be greater than or equal to $(100/k)D(X)$, so that the dominant portion $D(X)$ of the sensitive cell X will receive proportionately at least as much protection from above as does the corresponding $D(Y)$ for a cell Y on the dominance threshold.

Lower estimates of $D(Y)$ or $D(X)$ are obtained in a much more complex manner. As $D(X)$, considered as a cell (although in general it is not a tabulation cell), is the aggregate response of n or fewer respondents, then $D(X)$ and any subcell of $D(X)$ is sensitive and thus suppressed. Therefore, lower estimates of $D(X)$ are obtainable only through lower estimates of the corresponding $V(X)$, in the following manner. If a lower estimate n' of n and an upper estimate t' of the number of respondents t to a non-sensitive and published cell Y are known, then one may conclude $D(Y) \geq (n'/t')V(Y)$. In many publications, t is published or may be straightforwardly inferred from published data regardless of whether $V(Y)$ is published or not. As a result, analysis of the t's corresponding to published and suppressed cells would most certainly lead a serious data analyst to a precise determination of the value of the parameter n. Therefore, under the assumption that n and t are precisely known, the relative equivocation from below afforded $D(Y)$ by publishing $V(Y)$ for a non-sensitive cell Y equals $(D(Y) - (n/t)V(Y))/D(Y)$. For Y on the sensitivity threshold $D(Y)=(k/100)V(Y)$, this relative equivocation from below equals $1 - (n/t)(100/k)$.

Remark. For published cells Y, other lower estimates of $D(Y)$ may be obtained from $V(Y)$. However, under the mild restriction $(k/100) \geq n/t$ (recall that $t \geq n+1$ for non-sensitive Y), the lower estimate $D(Y) \geq (n/t)V(Y)$ was best possible among those considered.

If X is sensitive so that $V(X)$ is suppressed, then lower estimates $L(D)$ of $D(X)$ may be obtained from lower estimates $L(X)$ of $V(X)$ provided a lower estimate k' of k is known. As X is sensitive by assumption, then $D(X) \geq (k/100)V(X) \geq (k'/100)V(X)$. As $L(X)$ is a lower estimate of $V(X)$, then $V(X) \geq L(X)$ and hence $D(X) \geq (k'/100)L(X) = L(D)$.

To provide at least the same relative equivocation from below to $D(X)$ for sensitive X as to $D(Y)$ for Y on the sensitivity threshold, we define an acceptable lower estimate of $V(X)$ for a sensitive cell X to be any lower estimate which is less than or equal to

$$\left\{ \begin{array}{ll} 0 & t \leq n \\ (n/t)(100/k)^2 D(X), & t > n \end{array} \right\}$$

Remark. It would be useful to determine upper and lower sensitivity measures S^+ and S^- for which $S^-(X)$ and $S^+(X)$ measure the amount of additional suppression necessary to protect $D(X)$ from above and below, respectively. Theoretical and practical considerations indicate the desirability of requiring these measures to be subadditive and superadditive, respectively, as the following inequalities demonstrate. If X is sensitive and Y is a candidate cell for complementary suppression, then the union XUY will be non-sensitive if $S^-(XUY) \leq V(XUY)$; and a lower estimate $L(XUY)$ of the union XUY will be acceptable if $L(XUY) \leq S^-(X) + S^-(Y) \leq S^-(XUY)$. We may construct a subadditive function on the set of cells by assigning to each cell Y the minimum acceptable upper estimate of its corresponding $D(Y)$ i.e., by defining $S^+(Y) = (100/k)D(Y)$. However, the corresponding function which assigns to each cell Y the maximum acceptable lower estimate of its corresponding $D(Y)$ as determined above is not a subadditive or superadditive function. In terms of defining a sensitivity measure in the sense of [4], it would be desirable to determine a superadditive minorant $S^-(Y)$ of this function.

THE PUBLICATION NETWORK AND LOGICAL TABLES

By the term tabulation cell we shall mean any cell whose value for a particular statistic is either tabulated for publication or, although not explicitly tabulated, may be determined from the values of tabulated cells by linear techniques; and the term publication network shall

denote the set of all tabulation cells together with the collection of all linear relationships between them. A publication network is definable in terms of one or more independent parameters, such as membership in certain of several geographic sets, industry groups or industry types.

The publication network may be realized as a directed linear graph representing set-subset relationships between classes of tabulation cells. These set-subset relationships and the linear relationships between the tabulation cells mutually define each other. Each point on the directed graph corresponds to a class of tabulation cells and each directed line segment between graph points (nodes, vertices) corresponds to a set of linear equations between the members of the corresponding classes of tabulation cells. For example, the four geographic parameters United States, State, County and City-within-County are related hierarchically, so that the graphical representation of an associated publication network would consist of four points arranged vertically in the order above, with directed line segments from the points corresponding to United States to State, State to County and County to City-within-County. As each graph point has at most one superior on the graph, then this network is one-dimensional. A two-dimensional network would result if these geographically defined cells were further disaggregated by another strictly hierarchical set of parameters. For example, if, as in the U.S. Census of Manufactures, the responding universe comprises all manufacturing establishments, each classified according to geographic location of place of business and industry type (by 6-digit within 4-digit within 3-digit within 2-digit Standard Industry Code), then the publication network would be two-dimensional. The corresponding directed graph would consist of the four points of the strictly geographic graph previously mentioned, together with the sixteen possible combinations of each of four geographic types with the four industry types, with corresponding directed line segments between these 20 points.

As the maximum number of directed segments terminating at any graph point in the preceding example equals two, the publication network is two-dimensional. For example, the graph point corresponding to County by 3-digit industry type has precisely two directed segments terminating at it, one emanating from each of the graph points County by 2-digit industry type and State by 3-digit industry type. Each of these directed segments represents a class of linear equations, namely those equations between a specific county by a specific 2-digit industry type and this county by the 3-digit industry types which make up the given 2-digit industry type, and those equations between the state containing the county by one of those 3-digit industry types and all counties within this state by this particular 3-digit industry type. These two classes of linear equations may be brought together to form a class of two-dimensional statistical tables, each table of which displays the two-way disaggregation of a particular state by a specific 2-digit industry type for a given statistic by means of the counties within the state and the 3-digit industry types which make up the particular 2-digit industry type. This situation admits a straightforward generalization, subject to the following definition. A tabulation cell in a statistical table is an internal cell if it is not a marginal total or partial marginal total (i.e., cannot be disaggregated by subsets) in the table.

General Observation. Given a publication network and its associated directed graph, the tabulation cells and the linear relationships between these which define the publication network may be organized for each statistic into tables so that each tabulation cell appears as an internal cell in precisely one such table. Moreover, the dimension of this table is less than or equal to the number of directed segments terminating at the graph point corresponding to the tabulation cell.

One dimension of each of these tables represents the disaggregation of a tabulation cell corresponding to a superior graph point of the given interior graph point by the tabulation cells it comprises at the inferior graph point. For example, a state is broken down by its counties or a particular 2-digit industry group is broken down by its 3-digit industry groups as in the previously mentioned example. These tables may be constructed inductively from the "top" (the maximal points) of the graph downwards, and shall be referred to as the logical tables of the publication network. This definition is motivated in part to distinguish the logical tables from other tabular displays of the data. The importance of the logical tables become clear when the suppression problem is viewed globally, i.e., in the context of the entire publication network.

An ideal global solution to the suppression problem in a publication network may be described as follows. Associate a variable to each suppressed tabulation cell in the publication network and associate to each unsuppressed tabulation cell its value. These variables and constants are substituted into the linear equations defining the publication network. The publication network is thus realized as a system of linear equations. Through application of linear programming techniques, best-possible upper and lower estimates of the values of suppressed sensitive cells and sensitive unions of suppressed cells are obtained to yield best-possible interval estimates of the values of these cells. (Sensitive unions of suppressed cells are formed under dominance criteria within a linear relationship between sensitive and nonsensitive cells may be derived in which the largest n respondents dominate. Since the linear equation corresponding to this cell union is derivable, then the value of the cell union is effectively published). If the interval estimate thus obtained for any suppressed sensitive cell is unacceptable, then, according to an established suppression methodology, additional cells are suppressed (i.e., additional variables are introduced into the system) until no unacceptable interval estimates of suppressed sensitive cells may be obtained within the network. This suppression methodology must also be

sensitive to predetermined rankings of cells as candidates for complementary suppression, to historical precedent and to relevant policy to the extent that attention to these does not diminish the information content of the publications in disproportionate measure to their importance. Above all, this methodology should minimize over-suppression of cells so that as few cells of the smallest possible value be suppressed complementarily in the network.

Unfortunately, the computational enormity of the process just described renders this process virtually impossible to implement in all but the smallest and simplest (e.g. strictly hierarchical) poublication networks. To render the problem tractable in general (for example, in censuses or large surveys), the problem must be organized into a set of local problems for which valid local techniques can be developed, together with controls for maintaining consistency between these local analyses. The General Observation previously stated provides this organization.

As previously described, the network is organized into collections of logical tables for which each tabulation cell appears as an internal cell in precisely one logical table. Beginning with the logical tables formed at the maximal points on the directed graph and proceeding downwards through the graph (with respect to the partial ordering of the graph points imposed by the directed line segments), the logical tables are subjected to an intra-table disclosure analysis which performs complementary suppressions if necessary in each logical table until each incoming suppressed cell can only be acceptably estimated within the logical table. (The algorithmics of such intra-table techniques will be discussed in the next section). As each logical table completes disclosure processing, best-possible interval estimates of all suppressed cells are computed and acceptable interval estimates of the value of each complementary suppression created within this logical table are defined in terms of the relationship between such estimates and interval estimates of the values of the suppressed sensitive cells in the logical table. The acceptable interval estimates of the complementary suppressions thus defined are passed to any subsequently processed logical table in which the complementary suppression appears as a marginal total. This is done to insure that only acceptable estimates made be made of suppressed sensitive cells within the network. As each tabulation cell appears as an internal cell precisely one logical table, this processing sequence can be completed in one pass (i.e., without "backtracking" to reprocess a particular table) if the operant sensitivity criterion and the defined acceptable estimates resulting make it possible to adequately protect any sensitive cell in a logical table by suppressing only internal cells in the table. In general, to control the disclosure analysis and suppression process theoretically and operationally and to minimize over-suppression, it is advisable to adopt a suppression methodology which suppresses cells on the margins in logical tables only when no combination of suppressed internal cells within the table will suffice to protect the table's sensitive cells.

INTRA-TABLE COMPLEMENTARY SUPPRESSION METHODOLOGY

The problem of intra-table disclosure advoidance and complementary suppression methodology in a publication network is to adequately protect all cells and unions of cells which have been designated as suppressed in a logical table through the process of complementary suppression, while minimizing the adverse impact of this process on the quality of the published data. It is therefore necessary that adequate upper and lower levels of protection for these suppressed cells and that the adequacy of individual unsuppressed cells as complementary suppression candidates can be determined. Our major assumption is that the quality of the published data is adversely affected more by the suppression of a larger number of cells than by the suppression of fewer cells of perhaps larger aggregate value. This assumption is justified in a large publication network by the cascading effect of cell suppression, i.e., suppresisons at higher levels in the network force, in an unpredictable manner, more suppressions at lower levels in the network. Therefore, although in particular cases it may seem that the quality of the data is least affected by the suppression of many small cells in favor of suppressing a few large ones, the fact that each of these complementary suppressions must be protected at lower levels of the network and may force the suppression of large cells at lower levels indicates that suppressing fewer cells is the better strategy in general.

This strategy may be mitigated by preassigning a Prefer (for suppression) or Disallow (from suppression) status to individual suppression candidates prior to the intra-table analysis. These assignments should be respected unless they serve to render the intra-table problem intractable, in which case they must be selectively relaxed or ignored.

The objectives of study in intra-table complementary disclosure analysis are unions and differences of suppressed cells for which the value of the cell union or difference is effectively published (i.e., can be obtained from the values of published cells by linear techniques). As each complementary suppression is performed in the table in turn, the set of unions and differences of suppressed cells is changed. When this set is such that the value of none of its members may be derived as an unacceptable upper or lower estimate of the value of a sensitive or other suppressed cell, the intra-table analysis and complementary suppression process is complete for this logical table. A suppression methodology must be developed for which this sequence of complementary suppression terminates in a minimum or near-minimum number of complementary suppressions. This problem is significantly more difficult in three and higher dimensional logical tables than it is in one or two dimensions. Although operational programs based upon heuristic algorithms are being developed to

complementary suppression in three and higher dimensional tables, the subsequent discussion will be limited to the two dimensional case (of which the one dimensional case is a particular application). This limitation does not, however, apply to the techniques of linear estimation employed, which easily generalize to higher dimensional problems.

Although upper estimates of suppressed cells in a two-dimensional logical table can be obtained from the linear equations corresponding to the row and column containing the suppressed cell (i.e., the cell is estimated from above by the difference between the row or column marginal total, if it is published, and the sum of all published cells on the row or column), it is the set of all linear combinations of these line estimates which comprise all linear estimates of the value of the suppressed cells obtainable from the logical table. By means of these linear combinations, better upper estimates and nontrivial (i.e., positive) lower estimates of the values of suppressed internal cells in a logical table may be obtained. Techniques for obtaining such estimates are described in [1]. The problem of obtaining best-possible upper and lower estimates of cells in a logical table may be posed as a generalized transportation problem as studied in the field of operations research.

In the classical transportation problem, there are q supply points each with fixed supply and p demand points each with fixed demand. There is a transportation cost per unit delivered associated with each supply point-demand point association. Assuming total supply equals total demand, the transportation problem is to assign supply to demand so that the total transportation cost (the cost function) is minimized. The problem is represented by a $(p \times 1)^x$ $(q \times 1)$ array. The i-th row of this array corresponds to the i-th demand point, $1 \leq i \leq p$, the j-th column of the array corresponds to the j-th supply point, $1 \leq j \leq q$, the entry in position (i.j) is a variable x_{ij} representing the amount supplied by the j-th supply point to the i-th demand point, while the entries (i, q + 1) and (p + 1, j) are, respectively, the total demand at the i-th demand point and the total supply at the j-th supply point, $1 \leq i \leq p$ and $1 \leq j \leq q$. The entry (p + 1, q + 1) equals the common value of total supply and total demand. The reader is referred to [3] for a discussion of various classes of transportation problems and their solutions. In the disclosure application, each published cell in the logical table is replaced by its value. Unlike the classical transportation problem, some of the row and column marginal totals may be variables. The costs associated with each variable in the cost equation are taken from the discrete set {- 1, 0, 1}, so that, for example, if we seek to determine the minimum value (i.e., the best lower estimate) of the cell in the (1, 1) position, we minimize the cost function x_{11}. If we seek the maximum value of this cell (i.e., its best upper estimate), we find the minimum value of the cost function $-x_{11}$, and take its negative. Optimal estimates of cell unions and differences may be obtained by minimizing or maximizing the analagous linear relationships between their corresponding variables. Standard transportation problem techniques may be employed to determine these minima and maxima. The significant computational difference between this application and the classical transportation problem is that several iterations of the techniques may be necessary in the disclosure application before a feasible solution to the problem is reached (see [3]).

In [1], the author describes techniques for determining interval estimates of the values of suppressed cells in a logical table using an algorithm tailored to the disclosure problem. This algorithm begins with a line estimate of a particular cell or cell union and systematically generates cell unions and differences related to this cell, comparing the upper and lower estimates of the cell value thus obtained with previously obtained estimates. The algorithm operates quite efficiently and has never failed to obtain best-possible estimates. It remains to prove that this algorithm always generates best-possible interval estimates of the values of suppressed cells in a logical table (e.g., that this algorithm is equivalent to existing transportation algorithms). This is under investigation.

Although methods for determining best-possible interval estimates have been established, an area of research which remains open is that of determining a minimal set of complementary suppressions given a set of specified suppressions and their acceptable upper and lower estimates. An exhaustive examination of the alternative complementary suppression patterns is out of the question in all but the smallest logical tables; and no acceptable branch and bound procedure has yet been devised, although these remain under investigation. A geometric approach to the problem is indicated to provide guidance and control in the choice of complementary suppressions. Geometrically, we may view the disclosure problem as represented by a 0 - 1 matrix in which the position corresponding to a suppressed cell or a cell disallowed as a complementary suppression candidate contains a 0 and those corresponding to candidates for suppression contain a 1. For the moment ignoring the cell values and assuming that any one candidate for complementary suppression in a row or column will suffice to protect that row or column (i.e., the union of this cell with all suppressed cells on the row or column is non-sensitive), then a partial geometric solution of the suppression problem is provided by the following theorem.

Theorem. Let R and C denote the number of rows and columns, respectively, in a logical table which require additional suppression (the unprotected rows and columns). Assume that one additional suppression in an unprotected row or column will suffice to protect this row or column. If R=C=1, then at most three additional suppressions are necessary in the logical table to protect all rows and columns. Otherwise, Max (R,C) additional suppressions suffice. Assume for definiteness that R=Max (R,C). Then the

first C of these complementary suppressions must be chosen so that one is in each of the C unprotected columns and each is in a different row. The remaining R-C complementary suppressions are chosen with one in each of the remaining unprotected rows and each may be chosen in any column, provided that, if one is chosen in a column not containing any suppressions, then at least one other is chosen in the same column.

It results that the number of such solutions grows like the factorial of Max (R,C), so that many alternative suppression patterns exist. This theorem, when applied in conjunction with specified Prefer and Disallow suppression options and branch and bound techniques has proven effective in determining optimal or near-optimal suppression patterns which protect cells in their rows and columns in real disclosure settings (i.e., where one complementary suppression on a row or column may not suffice to protect the row or column, and where n respondent dominance in cell unions is a factor). If, after the Theorem has been applied, improved estimates of suppressed cells are obtained through linear combinations of row and column equations (i.e., from cell unions or differences which are formed through linear combinations of rows and columns), the suppression pattern generated by application of the Theorem is appropriately augmented. A generalization of the Theorem which identifies all single variable linear equations obtainable from a given suppression pattern and the corresponding set of covering suppressions is under investigation.

THE SYSTEM AS IMPLEMENTED

An automated system to perform disclosure analysis and complementary suppression for the 1977 Economic Censuses of Manufactures, Construction Industries and Wholesale and Retail Trade is currently completing development at the U.S. Bureau of the Census. This system is written in Fortran and its initial implementation will be on Univac 1100-series computers. The system applies the methodology described in this paper, with the following important exception.

There are four parameters employed to define the statistical cells in these publications, of which as many as three may be cross-tabulated to define a particular tabulation cell. These parameters are Geography, Standard Industry Code, Sales Type and Type of Establishment. The latter three of these are strictly hierarchical (i.e., one-dimensional), but the geographic parameter, owing to overlapping geographic regions, is two-dimensional. As almost all statistics are cross-tabulated by Geography, this implies that almost all logical tables will be at least three-dimensional. As no three or higher dimensional analog of the Theorem cited in the preceding section exists, it was decided to develop a methodologically sound two-dimensional complementary suppression computer program and to process only two-dimensional logical tables. This procedure is feasible in three-dimensional publication networks for which Geography is a defining parameter because data for overlapping portions of geographic regions are not published. Therefore, the corresponding cells may be employed as available suppressions, so that problems in the third dimension may be made to occur infrequently, and the constituent two-dimensional tables may be processed separately. When problems in the third dimension do occur, the processing order is backtracked in a well-defined manner.

The only four-dimensional tables constructed are those of Geography by SIC by Sales Type. Owing to the backtrack technique previously described, these four-dimensional tables can be regarded as sets of three-dimensional tables of one geographic dimension by SIC by Sales Type. To process these three-dimensional tables, each three-dimensional table will be partitioned into a collection of two-dimensional tables, one for each Sales Type. These will be processed separately by the two-dimensional suppression program. At various stages in this analysis, the collection of two-dimensional tables comprised by the original three-dimensional table will undergo a three-dimensional disclosure analysis reconciliation.

BIBLIOGRAPHY

[1] Cox, L., Statistical Disclosure in Publication Hierarchies, 1976 Proceedings of the Statistical Computing Section - American Statistical Association, pp. 130-136.

[2] Interim Report on Statistical Disclosure and Disclosure-Avoidance Techniques, Subcommittee on Disclosure-Avoidance Techniques, Federal Committee on Statistical Methodology, Statistical Policy Division, Office of Management and Budget, 1977. (unpublished)

[3] Dantzig, G., Linear Programming and Extensions, Princeton University Press, Princeton, 1963.

[4] Sande, G., Towards Automated Disclosure Analysis for Enterprise Based Statistics, Statistics Canada, 1977. (unpublished)