Suppression Methodology and Statistical Disclosure Control

Lawrence H. Cox

Journal of the American Statistical Association, Vol. 75, No. 370 (Jun., 1980), 377-385.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28198006%2975%3A370%3C377%3ASMASDC%3E2.0.CO%3B2-I

Journal of the American Statistical Association is currently published by American Statistical Association.

Suppression Methodology and Statistical Disclosure Control

LAWRENCE H. COX*

This article discusses theory and method of complementary cell suppression and related topics in statistical disclosure control. Emphasis is placed on the development of methods that are theoretically broad but also practical to implement. The approach draws from areas of discrete mathematics and linear optimization theory.

KEY WORDS: Statistical disclosure control; Complementary cell suppression; Linear sensitivity measure; Linear estimation and validation; Oversuppression.

1. THE DISCLOSURE PROBLEM AND CELL SUPPRESSION

Techniques of statistical disclosure control are aimed at protecting the confidentiality of individual respondents to a set of statistical publications. Effective disclosure control techniques reduce to an acceptable level the likelihood that either a respondent may be identified through its responses or that data collected from an identifiable respondent may be determined or narrowly estimated from the published data. Typically, these two problems are associated with *categorical* (or qualitative) data and *magnitude* (or quantitative) data, respectively. In this article, we consider only current data published in aggregate form. In particular, we do not discuss the confidentiality problems associated with time series data or microdata.

Disclosure in both categorical and magnitude data arises from unacceptably narrow estimation of the values of statistical cells. In the case of categorical data, a respondent may be identified if the value of a cell comprising all respondents that share certain attributes with the given respondent is sufficiently small and is published. If as a result, a confidential or sensitive characteristic such as income may be correctly attributed to an identified respondent, then *statistical disclosure* has occurred. Analogously, for magnitude data such as establishment-based economic data, if data such as total sales receipts for an identifiable respondent such as a large department store chain in a city can be narrowly estimated from published cell data, then disclosure results.

The fundamental objects of study in our analyses are the statistical cells for which the data are aggregated and their unions and differences, referred to as the *tabulation cells*. Without exception, we assume that the contribution of each respondent to a cell and hence the aggregate cell value is nonnegative.

Section 2 describes how in a census or major survey the typically large number of tabulation cells and linear relations between them necessitate partitioning a single disclosure problem into a well-defined sequence of interrelated subproblems. The approach is constructive and yields efficient data structuring and computer processing methodologies suited to the disclosure problem.

We do not discuss the qualitative considerations by which statistical cells and data are deemed confidential or sensitive. Rather, based upon established notions of cell sensitivity, Section 3 discusses mathematical techniques for measuring cell sensitivity. These techniques have immediate implication for Section 4, which discusses the problem of verifying, in part through linear estimation, whether adequate disclosure protection is in fact provided to individual respondents. This is referred to as the *validation phase* of the disclosure control process.

In Section 5 complementary cell suppression is discussed. Under cell suppression methodology, all cells identified as disclosure cells (or *sensitive cells*) are suppressed from publication, together with a sufficient number of additional nonsensitive cells, called the *complementary suppressions*, to ensure that the values of only nonsensitive cells and cell combinations may be derived from the published cell data. For example, if a cell representing data for a particular county is sensitive, then a sufficient number of additional cells at the county level within the given state will be suppressed complementarily until no union of county cells containing this sensitive cell is sensitive. Analogously, subcells of this sensitive cell as well perhaps as other cells must be suppressed complementarily to ensure that a narrow lower estimate of the value of this sensitive cell cannot be derived. We discuss cell suppression last to illustrate its dependency upon established notions of cell sensitivity (Section 3) and techniques of linear estimation and validation (Section 4).

Cell suppression may be applied to both categorical and magnitude data. However, whereas alternative disclosure control techniques exist for categorical data, currently there is no general operational alternative to

* Lawrence H. Cox is Principal Researcher Mathematician for application mathematics, Statistical Research Division, U.S. Bureau of the Census, Suitland, MD 20233. The author expresses thanks to the referees and an associate editor for their help in revising this article.

cell suppression for magnitude data such as establishment-based economic data.

For background on statistical disclosure, the reader is referred to U.S. Department of Commerce (1978) and Cox and Sande (1979), both of which contain extensive bibliographies.

2. ORGANIZING THE DISCLOSURE PROBLEM

As described in Section 1, the conceptual framework established in this paper divides the problem of controlling statistical disclosure through cell suppression into three areas: (a) developing well-defined procedures for identifying disclosure cells and for establishing necessary conditions for adequate disclosure protection to exist, (b) employing optimization techniques to verify that these necessary conditions have been met and verifying sufficient conditions directly, and (c) developing theoretically sound and computationally efficient algorithms to perform complementary suppression. However, the actual process of implementing cell suppression and validation techniques for a large and complex publication such as a census or major survey introduces a fourth area of study, that of *partitioning* a problem that cannot be solved singly into a well-defined sequence of subproblems, each of which is tractable computationally. In this section we discuss this partitioning problem with reference to a continuing example of a large publication set of aggregate cell data, the U.S. Census of Manufactures. A brief description of this census will now be provided.

In the Census of Manufactures, data for several statistical items such as "total capital expenditures" are collected from each manufacturing establishment in the respondent universe. For each statistical item, aggregate statistical cells defined by two independent sets of attributes uniquely associated with each individual establishment are tabulated. These attributes are the geographic location of the manufacturing establishment and the principal type of manufacturing activity in which the establishment is engaged, designated by a multidigit Standard Industry Code (SIC). The geographic attribute is defined by unique association with one member in each of four lists of geographic identifiers, State, Standard Metropolitan Statistical Area (SMSA) (or non-SMSA), County, and Place lists. Each of these lists covers the entire United States. Although not true in general, the Place (city or rural area) in which the establishment is found is often contained entirely within its County, which in turn is often contained entirely within its SMSA, which in turn is often contained within a particular State, so that in these cases geographic relationships are hierarchical. Analogously, tabulations by SIC are made on the basis of membership in a set of establishments with the same first 6 digits in their SIC, the same first 4 digits, 3 digits, and 2 digits, so that SIC's are related hierarchically. To the set of four geographic identifiers is appended the universal identifier "United States," and to the set of four SIC identifiers is appended the identifier "zero-digit SIC," so that tabulations strictly by SIC or geography may be made as well. For example, tabulating the state "New York" by "zero-digit SIC" for the statistical item total capital expenditures produces the aggregate value of all establishments located in New York state for this item, regardless of individual manufacturing activity.

Tabulations are thereby defined for any statistical item and any pair (geography, SIC) such as the State $\times$ 3-digit SIC pair (New York, 217). Each such pair defines a *statistical cell* consisting of the set of all establishments located within the given geographic area and active in the given manufacturing activity. Some of these cells may be empty. The value of this cell for the statistical item equals the aggregate of the nonnegative values for this item over all establishments in the cell. Because the cells are related by set-subset relationships (e.g., the cell (United States, 217) equals the disjoint union of all cells (S_q, 217) over all states S_q), their values for a statistical item such as total capital expenditures are related by linear equations wherein a unique variable corresponds to the value of each cell for this item. Organizing the disclosure problem amounts to organizing this potentially vast collection **E** of linear equations.

Under cell suppression methodology, all sensitive cells are suppressed from publication, perhaps together with additional, nonsensitive cells sacrificed as complementary suppressions to obscure the values of the sensitive cells. In the most rigorous sense, complete disclosure control is achieved when, after replacing each variable of **E** corresponding to a published (i.e., unsuppressed) cell by its value, only linear combinations of the variables of **E** corresponding to nonsensitive cell combinations can be derived directly through manipulation of the linear equations of **E**. Techniques described in Section 3 permit computation of *minimal levels of protection* for sensitive cells and cell combinations, that is, necessary conditions that must be met in order for respondents in these cells to be afforded adequate disclosure protection. The minimal levels of protection establish minimum upper and maximum lower estimates of the values of these cells that will be tolerated.

In the broadest throretical sense, a *global solution* to the disclosure problem must provide optimal interval estimates $[\min(v), \max(v)]$ of the values of the variables v of **E** corresponding to the sensitive cells, as well as of linear combinations of the variables corresponding to potentially sensitive cell combinations, and must verify that all such derived interval estimates are tolerable (or *acceptable*). As we shall see in Section 3, these conditions are unfortunately not sufficient to guarantee nondisclosure. Once these necessary conditions have been met, potentially sensitive cell unions must be individually examined to assure that complete disclosure protection has been provided to all respondents. This latter question of sufficiency can lead to significant practical problems discussed in Section 4.

The preceding global solution to the disclosure problem has two principal drawbacks which render it infeasible or inadequate in general. First, if the number of variables in **E** is large (as is certainly the case in a census or major survey), it may be impractical or impossible to subject **E** to many passes through a general-purpose linear program or a technical equivalent to obtain the required interval estimates. Second, and more important theoretically, **E** has not been equipped with any additional structure that would indicate good or best choices for the complementary suppressions. Without such structure to guide the complementary suppression process, *oversuppression* of cell data could not be minimized, and, in an automated system, processing efficiency could not be guaranteed. For large problems, therefore, **E** must be structured into a collection of interrelated *local problems*. We shall briefly describe how this may be accomplished.

As all linear equations e in the collection **E** represent disaggregation relationships, then each is of the form

$$e: v = v_1 + v_2 + \ldots + v_t \,, \tag{2.1}$$

where v and the v_i are the variables uniquely assigned to the values of particular cells for a given statistical item. For example, v could represent the aggregate value of total capital expenditures in the cell (United States, 217), and each v_i could represent total capital expenditures in the cell (S_q, 217). To avoid processing ambiguities, we require that all intermediate aggregations be explicitly represented. To do so, it suffices to assume that all linear equations e of **E** are *reduced*, that is, no linear equation f: $w = v_{i1} + \ldots + v_{im}$ of **E** exists for which $\{v_{i1}, \ldots, v_{im}\}$ is a subset of $\{v_1, \ldots, v_t\}$ for $m \geq 2$. Each variable v of **E** that occurs on the left side of (2.1) is called an *aggregate variable* of **E**.

The linear equation e may be viewed as a one-dimensional tabular array representing the disaggregation of v. If v is also disaggregated in a second equation of e' of **E**

$$e': v = v'_1 + v'_2 + \ldots + v'_s \tag{2.2}$$

(as would be the case for v' equal to total capital expenditures in the cell (United States, 217)) and each v'_j is equal to total capital expenditures in the cell ((United States, 217k), where 217k represents a 4-digit sub-SIC of the 3-digit SIC 217), and if the v_i and v'_j are also aggregate variables of **E**, then both of these disaggregation relationships may be brought together in the form of a single two-dimensional tabular array **T** as illustrated by Table 1. In Table 1, each row of **T** corresponds to an equation of **E** that disaggregates one v'_j; and each column of **T** corresponds to an equation of **E** that disaggregates a v_i. In our ongoing illustration, v''_{kq} equals the aggregate value of total capital expenditures in the cell (S_q, 217k). If v were disaggregated in a third equation of **E** (a situation that does *not* occur in our illustration), or a fourth, ..., or mth equation, then three, or four, or ..., or m-dimensional tabular arrays result. We call these derived tabular arrays the *logical tables* of **E**. There are precisely as many logical tables as there are aggregate variables of **E**. As each equation of **E** is represented in one logical table, then the collection of logical tables together with the respondent data contain all information necessary to perform a complete complementary suppression methodology on **E**. In particular, we need not consider any other tabular displays of the equations and variables of **E**.

1. A Logical Table

v_{11}''	$v_{12}'' \cdots v_{1t}''$	v_1'
v_{21}''	$v_{22}'' \cdots v_{2t}''$	v_2'
⋮	⋮	⋮
v_{s1}''	$v_{s2}'' \cdots v_{st}''$	v_s'
v_1	$v_2 \cdots v_t$	v

As implied by the results of the next section, the proper choice of a cell sensitivity criterion assures that if in the linear equation (2.1) one of the disaggregates v_i corresponds to the value of a sensitive cell, then the aggregate variable v, if nonsensitive, need not be suppressed to protect v_i from narrow estimation from above in this equation. This condition implies that oversuppression can be minimized and processing efficiency maintained if the cell suppression and validation processes are first performed on the highest level aggregations (i.e., those at the United States and 0-digit SIC levels) and successively on lower level aggregations, in an order that demands that no logical table representing the disaggregation of one of the v_i from (2.1) has been processed before the logical table corresponding to v. Anomalous patterns of cell values may produce finite cycles in this processing sequence. However, because each variable of **E** appears as an *internal entry* in at most one logical table, in general the processing sequence is not disrupted operationally.

We conclude this section by formalizing these results in the language of linear graph theory. This formulation lends itself to the design and implementation of efficient data structures and a computer processing methodology to solve the disclosure problem operationally.

A (finite) *directed linear graph* **G** consists of a set $\mathbf{P} = (p_1, \ldots, p_r)$, comprising the *nodes* of **G**, together with a collection of ordered pairs (p_i, p_j) of distinct members of **P**, called the *arcs* of **G**. The nodeset of the graph **G** associated with the linear system **E** is defined to be the set of all variables of **E**, each of which corresponds to a unique tabulation cell by definition. The arcset of **G** consists of all ordered pairs (v, v_i) for which v_i appears as a disaggregate of v in some equation e of **E** of the form (2.1).

Because each node except terminal nodes of **G** (i.e., the nodes corresponding to the nonaggregate variables of **E**) corresponds to a unique logical table **T** of **E**, then **G** provides a natural mechanism for representing the previously defined processing sequence of the logical tables. For the Census of Manufactures example, the first logical tables processed are the one-dimensional

logical tables representing statistical item totals at the United States level broken down by State and at the United States level broken down by its 2-digit SIC's. Next, each State total is broken down by 2-digit industry; then, each 2-digit industry total is broken down by State, and so on. Finally, each County by 4-digit industry cell total is broken down in a two-dimensional table consisting of Place by 6-digit industry internal entries.

This approach has been implemented in a data base environment in an automated disclosure control system for the 1977 U.S. Economic Censuses. We remark that beginning at the terminal nodes of the graph **G** and traversing the arcs in reverse order, one defines a "roll-up" tabulation system for the statistical cells that may also be straightforwardly implemented once appropriate data structures representing **G** have been constructed.

3. SENSITIVITY MEASURES AND CELL VALUE ESTIMATION

The preceding section dealt with organizing the tabulation cells for disclosure processing. Sections 4 and 5 discuss techniques of linear estimation and validation and complementary cell suppression. In order to effectively implement these techniques efficiently within a logical table, it is useful to first determine a lower bound on the amount of complementary suppression necessary to render the logical table disclosure-free. As we subsequently demonstrate, such techniques for determining minimal levels of protection for the sensitive cells follow naturally from cell sensitivity criteria if these criteria are mathematically well defined. Although we use the term *cell* throughout this discussion and concentrate on the problem of obtaining cell value estimates from above, analogous techniques apply equally to cell combinations and problems of estimation from below.

Assuming that a notion of confidential or sensitive data has been established, we develop formalisms and techniques relative to which the degree of disclosure may be measured. For aggregate data, the degree of disclosure risk and threat to the confidentiality of respondents is measured and analyzed at the statistical cell level. We begin with examples of *cell sensitivity criteria* in current use by statistical agencies.

In frequency count tabulations of categorical data, a respondent contributes *1* to the cell value if the respondent possesses all attributes that delineate the cell, and contributes *0* otherwise. Cell sensitivity for categorical data is typically determined by a *threshold rule*, which, for an appropriately chosen integer n, defines a cell to be sensitive if and only if the cell contains n or fewer respondents. The fixed value n is chosen so that the probability of identifying any individual as a member of an arbitrary set of $n + 1$ or more respondents within the respondent universe is acceptably small with respect to established criteria.

For magnitude data such as establishment-based economic data, threshold criteria of cell sensitivity are no longer adequate. For example, a given cell may contain any number of respondents, but the contribution of one of these respondents may completely dominate the cell, in which case the cell value is an unacceptably narrow upper estimate of this respondent's contribution, and the cell is sensitive. Even if one respondent does not dominate the cell, the aggregate of the responses of a small subset of respondents may. As a result, members of this subset may individually or collectively employ knowledge of their own contribution to narrowly estimate that of another respondent, so that again the cell is sensitive.

For these reasons, a more general measure of cell sensitivity than threshold criteria is required for magnitude data. One such measure, employed in the U.S. Economic Censuses, is the *(n, k)-dominance rule*, which defines a cell to be sensitive for a particular statistical item if n or fewer respondents constitute greater than k percent of the total cell value for this item. As respondents in economic censuses and surveys are easily recognized by their geographic location(s) and type(s) of manufacturing or business activity, the value n is not chosen on the basis of respondent recognizability, but rather is chosen to be larger than the size of coalitions of respondents presumed to exist. Such coalitions pool their individual confidential data for the purpose of determining the confidential data of a competitor. For example, if a coalition of three respondents exists in a cell containing only four respondents, then the contribution of the remaining respondent may be precisely determined by the coalition.

As approximate disclosure of magnitude data is for most practical purposes no less damaging than exact disclosure, the value k is chosen so that unacceptably narrow estimates of single respondent data cannot be derived. Thus, in the above example the choice of a dominance rule with $n = 4$ and $k = 80$ would ensure that in publishing the value of a nonsensitive cell the contribution of the largest contributor in the cell would be equivocated by at least 25 percent (20/80) from the scrutiny of a coalition of three or fewer of its competitors.

Different assumptions regarding techniques available to the confidentiality invader produce different sensitivity criteria (Cox 1979). The level of sensitivity in a cell or cell combination is defined by the ambient cell sensitivity criterion. In the remainder of this section, we discuss techniques for representing sensitivity criteria through well-defined real-valued functions S defined on the set of all cells and cell unions over the data set. These functions S are called *sensitivity measures*. By definition, a cell X is sensitive if $S(X) > 0$, and is nonsensitive otherwise.

A sensitivity criterion should reflect the desired balance between respondent confidentiality and the data needs of the user community and therefore should not unnecessarily preclude the publication of data. In particular,

any sensitivity criterion under which the union of two nonsensitive (and therefore potentially publishable) cells fails to remain nonsensitive is certainly faulty. By representing a sensitivity criterion as a sensitivity measure S, the requirement that the union of two nonsensitive cells remain nonsensitive under S is expressible in terms of the mathematical notion of subadditivity.

Definition 3.1. A real-valued function S defined on the set of all tabulation cells in a data set is *subadditive* if for any tabulation cells X and Y and their cell union $X \cup Y$ the relation

$$S(X \cup Y) \leq S(X) + S(Y) \tag{3.1}$$

holds.

For example, the (n, k)-dominance rule is subadditive because the percentage contribution of a group of respondents in the union of two cells cannot exceed the maximum of their percentage contributions in either cell. Examining the (n, k)-rule further, if we denote the nonnegative contribution of the N respondents to the cell X by x_i, $1 \leq i \leq N$, and if we choose to order the respondents of X so that $x_1 \geq x_2 \ldots \geq x_N \geq 0$, then setting $x_i = 0$ for $i > N$, we observe that the (n, k)-rule may be stated in the following form.

(n, k)-dominance rule: If

$$S_D(X) = \sum_{i=1}^{n} ((100 - k)/100)x_i + \sum_{i=n+1}^{\infty} (-k/100)x_i , \tag{3.2}$$

then a cell X is sensitive under the (n, k)-rule if and only if $S_D(X) > 0$; that is, $S_D(X)$ is one sensitivity measure representing the (n, k)-rule. Any nonnegative multiple of $S_D(X)$ would also suffice. By considering $S_D(X)/V(X)$, where $V(X) = \sum_{i=1}^{\infty} x_i$ is the *value* of the cell X, we may measure the level of sensitivity of individual cells and compare these levels between different cells. $S_D(X)$ is a subadditive linear function of the x_i. We briefly sketch results in the study of sensitivity measures.

Definition 3.2. A *linear sensitivity measure* S is a real-valued function defined on the set of all tabulation cells in a data set and has the form

$$S(X) = \sum_{i=1}^{\infty} w_i x_i \tag{3.3}$$

wherein the w_i are constants, and $0 \leq x_i \leq x_j$ whenever $i \geq j$.

Linear sensitivity measures were introduced in Sande (1977) and their mathematical properties were investigated in Cox (1979). The salient results of Cox (1979) are stated below without proof. These and related results are important in the development of sound disclosure definitions and complementary cell suppression strategies. The first of these results provides a simple means of verifying whether a proposed sensitivity criterion passes the acid test of subadditivity.

Theorem 1: Let $S(X) = \sum_{i=1}^{\infty} w_i x_i$ be a linear sensitivity measure. $S(X)$ is subadditive if and only if $w_i \leq w_j$ whenever $i \geq j$.

The search for complementary suppressions could be shortened if only sufficiently large candidate cells need be examined. This assumes that an answer to the following question can be found: Given the sensitive cell X, how large must $V(Y)$ be in order to ensure that $S(X \cup Y) \leq 0$? This problem is somewhat more complex than it might appear, because in magnitude data such as establishment-based economic data, a single respondent such as a large conglomerate corporation may contribute to several cells. Its contribution to the union of these cells is by definition the aggregate of its contributions to the constituent cells. As a result, this single respondent may be the smallest contributor in each constituent cell but emerge as the largest contributor in the cell union. Its aggregate contribution to the cell union must also be kept confidential. The following theorem provides a partial answer to our question.

Theorem 2: Let $S(X) = \sum_{i=1}^{\infty} w_i x_i$ denote a subadditive linear sensitivity measure. Let X denote a sensitive cell and let Y denote an arbitrary cell. Assume that there exists an integer m such that $w_i = w_m$ for $i \geq m$ (a condition that holds for all sensitivity measures encountered in practical application). A necessary condition that the cell union $X \cup Y$ be nonsensitive is that the value $V(Y)$ of Y satisfy $V(Y) \geq S(X)/|w_m|$.

Unfortunately, this necessary condition is not sufficient if respondents can contribute to more than one cell in a cell union. Consider for example the (3, 75)-dominance rule together with two cells X and Y. Let the respondents to X be A, B, and C, and let their contributions be $A(70)$, $B(20)$, and $C(9)$, so that $V(X) = 99$ and $S(X) = 99/4$. Theorem 2 asserts that a cell Y, which adequately protects X from above, must satisfy $V(Y) \geq (99/4)/|-\frac{3}{4}| = 33$. If Y consists of respondents and contributions $D(9)$, $E(9)$, $F(8)$, and $G(7)$, then $V(Y) = 33$ and $S(X \cup Y) = 0$, so that $X \cup Y$ is nonsensitive. However, if Y consists of $H(15)$, $I(9)$, and $J(9)$, then $V(Y) = 33$ but $S(X \cup Y) = 6$, so that $X \cup Y$ is sensitive. In addition, were Y to consist of $A(15)$, $I(9)$, and $J(9)$, then the union $X \cup Y$ would consist of $A(85)$, $B(20)$, $C(9)$, $I(9)$, and $J(9)$, so that $S(X \cup Y) = 15$. The following theorem demonstrates that equally pessimistic results are true in general.

Theorem 3: Let S denote a subadditive linear sensitivity measure and let X denote a sensitive cell relative to S. Assume that respondents can contribute to more than one cell in a cell union. Then there does not exist a value $T(X)$ for which the condition $V(Y) \geq T(X)$ alone guarantees that the cell union $X \cup Y$ will be nonsensitive.

Theorem 1 provides a straightforward means for identifying faulty cell sensitivity criteria. Theorem 2, as a necessary condition for the completeness of the com-

plementary suppression process, provides a means of improving the efficiency of that process operationally. Theorem 3 demonstrates that in certain applications attention must be paid at the respondent level in complementary suppression to ensure that sensitive cell unions are not published. Technical tricks may be employed to reduce the seemingly enormous computational burden implied by Theorem 3 in the examination of all potentially sensitive cell unions. However, the computational requirements of these examinations may nonetheless be significant.

4. VALIDATION AND LINEAR ESTIMATION

An effective cell suppression methodology should integrate the processes of validation and linear cell value estimation and complementary cell suppression to greatest extent possible. This is because, ideally, the validation process should provide relevant information directly to the complementary cell suppression process. For this reason, we discuss validation and linear estimation first. Following the results of Section 2, we focus the discussion on the problem of linear estimation in two-way tabular arrays, the logical tables. Although we limit the discussion to techniques for determining upper estimates of the values of suppressed cells and cell combinations, analogous results are available for determining lower estimates of these values.

Effective techniques of linear estimation in two-way tabular arrays should exhibit three principal attributes. They should be *exact*, which means that they always produce optimal upper and lower estimates of the values of suppressed cells and cell combinations. They should also be computationally efficient. Finally, the validation process ideally should provide relevant information directly to the complementary suppression process. Such information typically consists of sets of potentially sensitive cell combinations from which good choices for the complementary suppressions can be made.

Validation amounts to assuring that only nonsensitive cells and cell combinations are published. A necessary condition for nondisclosure in cell unions is that all estimates of the values of sensitive cells and cell combinations be acceptable. This requires the determination of exact interval estimates of the values of all sensitive cells and potentially sensitive cell unions. In turn, this is equivalent to computing the maximum and minimum values attainable by each single variable v in the linear system **E** corresponding to a sensitive cell and attainable by each linear combination of variables corresponding to a potentially sensitive cell union. The problem of determining upper estimates, that is, the maximum values, in either case reduces to the problem of obtaining values of linear combinations of the variables of **E** that involve only the coefficient values 0 and +1.

For example, consider a suppressed sensitive cell X_1 and an arbitrary suppressed cell X, which appears in a logical table with X_1. Let v_1 and v_2 be their corresponding

2. A Table With Suppressions

x_{11}	x_{12}	x_{13}	32	185
10	x_{22}	x_{23}	72	107
x_{31}	4	67	x_{34}	174
x_{41}	54	89	x_{44}	222
131	150	223	184	688

variables in **E**. In the presence of a subadditive linear sensitivity measure, Theorem 2 may be applied to compute the quantity $U(X_1) = V(X_1) + S(X_1)/|w_m|$, which the value of any nonsensitive cell or cell union containing X_1 must attain or exceed. In order for adequate disclosure protection to be provided, it is necessary that $\max(v_1) \geq U(X_1)$ and, consequently, the maximum value of any $\{0, +1\}$-linear combination (or *aggregation*) of the v's involving v_1 (such as v_1+v_2) must also be greater than or equal to $U(X_1)$. These linear combinations may be *effectively published aggregations* (such as if X_1 and X_2 are the only suppressed cells on a row or column of a published table). Otherwise, the value of a linear combination such as $v_1 + v_2$ is bounded above by the value of some effectively published aggregation containing the expression $v_1 + v_2$. The importance of effectively published aggregations may be appreciated immediately by examining Table 2 in which the aggregation $x_{11} = 19$ is effectively published.

Two well-developed mathematical theories provide separate frameworks for studying the problem of determining maxima of $\{0, +1\}$-linear combinations of nonnegative variables, the theory of convex cones and the theory of transportation (or transshipment) problems. The reader unfamiliar with these theories is referred to Kuhn and Tucker (1956) and Dantzig (1963). We only sketch the relevant theories. The respective algorithms and examples may be found in Sande (1977) and Cox (1978b).

As Sande (1977) describes, the problem of obtaining upper estimates of the values of sensitive cells and cell combinations may be approached from the theory of convex cones. It is well known in this theory that any ray of a convex cone may be uniquely expressed as a nonnegative linear combination of the extreme rays of the cone (for our purposes, a ray of the cone corresponds to an effectively published aggregation). The extreme rays correspond to the *elementary aggregations* of the system **E**, that is, those effectively published aggregations not expressible as nonnegative linear combinations of other effectively published aggregations. For example, in Table 2, there are nine elementary aggregations: $x_{11} = 19$, $x_{12} + x_{13} = 134$, $x_{31} + x_{41} = 120$, and the second, third, and fourth row and column equations. Sande concluded that standard techniques from the theory of convex cones could be applied to compute the elementary aggregations of **E**. Upper bounds on the maxima of arbitrary aggregations of **E** corresponding to potentially sensitive cell unions could then be computed

from the fixed values of sets of elementary aggregations whose sum contains the given aggregation.

Subsequently, Cox (1978a) proved that this convex cone technique does in fact produce exact upper estimates of the values of arbitrary aggregations and demonstrated how in conjunction with techniques of interval arithmetic these techniques could also be applied to produce exact lower estimates as well. In addition, the convex cone technique can provide relevant information to the complementary suppression process in the form of the elementary and derived aggregations. Unfortunately, the convex cone technique does not appear to be at all computationally efficient, particularly since it must compute the maximum values of potentially many otherwise uninteresting aggregations in order to derive the minimum value of a single variable or aggregation.

The theory of transportation problems may also be applied to the disclosure problem. Within a single two-way tabular array, the linear relations between the variables corresponding to the suppressed cells consist of linear combinations of the row and column equations of the tabular array. For the special case of all internal cells of the array suppressed and all totals cells published, this collection of linear relations between the variables corresponds to a classical transportation (or transshipment) problem. In the more typical case in which not all internal cells need be suppressed and not all totals cells are necessarily published, the system of linear relations is equivalent to the linear constraint system for a transportation problem with zero-restricted variables. This formulation results from subtracting the value of each published internal cell from the value of its row and column cells, so that the published values may be considered to be "effectively zero" (see Cox 1977). Exact and computationally efficient algorithms exist for solving transportation problems. These algorithms could be exploited to provide the necessary information to the complementary suppression process. However, the efficiency of these specialized techniques deteriorates as the pattern of zero-restricted variables moves the linear estimation problem out of the realm of classical transportation problems and into the realm of general linear programming problems.

General purpose linear programming software will produce exact estimates of suppressed cell values and their combinations. In general, however, such software will be less efficient than specialized algorithms and does not provide any additional information to the complementary suppression process.

Cox (1978b) describes a modification of general linear programming theory tailored to the linear estimation problem in disclosure control. General linear programming theory is based upon the *simplex algorithm*. The simplex algorithm is a procedure for maximizing or minimizing a single linear combination of the variables, called the *objective function*, subject to linear constraints (in this case, the row and column equations of the tabular array). It does so by manipulating the linear constraints and the objective function through a sequence of operations called *pivots* until a desired form is achieved. To invoke the simplex algorithm twice (once for the maximum and once for the minimum) for each sensitive cell and potentially sensitive cell combination would probably be quite expensive. The algorithm of Cox (1978b), in contrast to general linear programming techniques, does not proceed objective function by objective function, but rather manipulates the linear constraints and observes at each pivot step which aggregations have been optimized by the current pivot operation. Limited but informative computational experience has demonstrated that this procedure results in the computation of all interval estimates in considerably fewer pivot steps than by general linear programming. For example, this algorithm computed the 18 single variables optima in Table 2 in 8 pivots, 6 of which were necessary to achieve an initial canonical form. In addition, the aggregations from which this specialized algorithm computes these optima are derived explicitly.

Theorem 2 may be applied to compute minimal levels of protection for sensitive cells and cell combinations. Techniques of linear estimation as described in this section may then be applied to verify that these necessary conditions for adequate disclosure protection have been met. As Theorem 3 indicates, however, these assurances are not always sufficient. When they are not, each effectively published cell union must be examined for sensitivity in order to guarantee that disclosure has been completely controlled. To do so directly, cell union by cell union, would probably be computationally infeasible. Indeed, to maintain and manipulate all attendant respondent data significantly reduces the computing resources available to perform other essential disclosure control functions. Even though technical mechanisms are available to reduce the number of cell unions to be examined, an unmanageable number of unions may still remain. For these reasons, in any specific data set the problem of disclosure in arbitrary unions should be carefully examined, perhaps probabilistically, so that a prudent balance between available resources and disclosure risk may be struck.

5. COMPLEMENTARY CELL SUPPRESSION

A discussion of complementary cell suppression is the final stage in our description of the disclosure control problem for magnitude data. From an unambiguous definition of disclosure (i.e., a subadditive sensitivity criterion) and exact techniques of linear estimation that explicitly produce effectively published aggregations, the data releaser seeks to develop efficient techniques of complementary cell suppression that employ this information to maximum advantage. To control disclosure from above, a data releaser seeks to develop techniques of complementary cell suppression that ensure that the values of only nonsensitive cell combinations are effectively published. Moreover, the releaser ideally wants to

minimize oversuppression of data, so that according to some predefined measure of information the amount of information suppressed from publication in the chosen suppression pattern is minimal in comparison with alternative suppression patterns.

Unfortunately, at its current stage of development the complementary cell suppression problem is not unambiguously defined. In general, notions of good or acceptable suppression patterns are based upon the completeness of the pattern (the condition that all effectively published aggregations are disclosure-free) and the degree to which the pattern is optimal with respect to minimizing oversuppression of data or information. As the question of completeness was discussed in the preceding sections, we restrict our present comments to the concept of oversuppression.

Given two complete suppression patterns containing the same sensitive cells as defined by an ambient subadditive linear sensitivity measure, several measurement functions are possible for determining the degree to which one pattern suppresses less data or information than the other. Among these measurement functions are (a) the total value (or the maximum or some other appropriate function) of the evaluation of the sensitivity measure on the elementary aggregations, (b) the total value of the suppressed cells in the pattern, (c) the number of respondents represented in the suppressed cells, (d) the total number of cells suppressed, and (e) some mixture of criteria (a)–(d). An *optimal pattern* is defined as one that achieves the optimal numerical value for the chosen measurement function. *Oversuppression* is then defined as the choice of any inoptimal suppression pattern. The choice of the measurement function by the data releaser will in general depend upon the data, its users and anticipated uses, and is by no means restricted to criteria (a)–(e). However, in applications for which compelling reasons to the contrary do not exist, it is reasonable to define optimal patterns by a procedure (e) for which the primary selection criterion is (d), the minimization of the total number of suppressed cells. This follows from two principal considerations.

In an application that has been partitioned into a set of local subproblems, suppressions cascade "downwards" through the directed graph from logical tables processed earlier to logical tables processed later in the processing sequence. For all practical purposes, it is impossible to predict the impact that suppressions made at earlier stages will have on the resulting suppression requirements at later stages. Under criteria other than (d), there is nothing to prevent achieving a global suppression pattern that is highly inoptimal with respect to the measurement function and involving many unnecessary suppressions by these means, even though all local solutions are optimal. By establishing (d) as the primary measurement criterion, a greater degree of process control and minimization of oversuppression should be maintained. In addition, criterion (d) allows for a concise and unambiguous statement of local and global optimality in mathematical terms. Analogous to the definition of sensitivity measures, this condition is necessary to develop well-defined and efficient solution strategies.

The problem of determining the minimum number of complementary suppressions needed to adequately protect the sensitive cells is unsolved for three and higher dimensional arrays and is only partially solved in two dimensions, even under appropriate limiting assumptions. This to some extent may be due to the fact that the corresponding techniques of linear estimation and construction of effectively published aggregations are themselves new. Although heuristic complementary suppression algorithms have been effectively applied in higher dimensions, the following discussion is restricted to the problem of determining the minimum number of complementary suppressions necessary to adequately protect the sensitive cells in arbitrary two-dimensional tabular arrays. This is referred to as the *two-dimensional complementary suppression problem.*

It is clear that no closed-form solution to the two-dimensional complementary suppression problem can exist without further limiting assumptions. This is because the pattern of the sensitive cells, their values and respondent contributions, and the values of the candidate cells for complementary suppression may be arbitrarily chosen, which implies that the optimal solution in any particular table can only be determined through an exhaustive search or branch-and-bound procedure. There is, however, a common mathematical thread that runs through all two-dimensional problems. It is in terms of this common thread that we define the major combinatorial problem underlying the two-dimensional complementary suppression problem.

The salient mathematical property common to all complementary suppression problems is that a necessary condition for the complementary suppression process to be complete is that there exist at least two variables in every effectively published aggregation involving the value of a sensitive cell. In the two-dimensional case, the linear space of all such aggregations is generated by the row and column equations in the two-dimensional table. Therefore, the combinatorial problem underlying the two-dimensional complementary suppression problem is that of determining, for an arbitrary pattern of suppressed cells in a two-dimensional tabular array, the minimum number of additional suppressed entries necessary to ensure that every effectively published aggregation in the augmented linear system contains at least two variables; and to exhibit at least one such minimal complementary suppression pattern. In addition, it is desirable to describe an algorithm for generating all such minimal patterns. This latter requirement is of particular practical importance because a solution to our combinatorial problem need not necessarily be a complete solution to the practical disclosure problem. It may therefore be necessary to compute all such minimal patterns to find one, if any at all, that is sufficient. What is less discouraging, however, is that the minimal patterns are frequently

most helpful in determining an optimal or adequate solution to the practical disclosure problem for a given table.

The combinatorial problem itself has not yet been completely solved. We present a theorem from Cox (1975) that partially solves the combinatorial problem. This theorem identifies all minimal complementary suppression patterns that complete a given pattern of suppressed cells within their row and column equations, that is, it identifies all minimal patterns with the property that there are at least two suppressed entries on each row and column. This result does not completely solve the problem, however, as it may be that a single-variable linear equation may be derived as a linear combination of these row and column equations. For example, in Table 2 the conclusion $x_{11} = 19$ follows from manipulation of the row and column equations, even though each row and column equation contains at least two suppressed entries (subtract the sum of the second and third column equations from the sum of the first and second row equations). Despite these such shortcomings, this theorem has proven effective in generating suppression patterns that are either minimal or that can be augmented to produce minimal patterns when applied in conjunction with linear estimation techniques, particularly because of the many minimal patterns it typically constructs.

Theorem 4: Let R and C denote, respectively, the number of rows and columns of a two-dimensional tabular array that contains precisely one suppressed cell (the unprotected rows and columns of the table). If $R = C = 1$, then the suppression pattern may be augmented so that each row and column of the tabular array contains at least two suppressed cells by the suppression of at most three additional cells. Otherwise, the pattern may be so augmented by suppressing precisely $\max\{R, C\}$ additional cells. Moreover, the number of such "minimal" patterns is a combinatorial function that often grows like $\max\{R, C\}$ factorial.

Because the proof of the theorem is constructive, it provides an algorithm for obtaining such "minimal" patterns. To sketch the proof in the general case, assume that $R = \max\{R, C\}$ and that $R > 1$. Choose the first C additional suppressions each in a different unprotected column of the table in such a manner that each is also in a different unprotected row of the table, thereby protecting all columns of the table and R additional rows of the table. The remaining $(R - C)$ additional suppressions are chosen to be in the remaining $(R - C)$ unprotected rows of the table, and each may be chosen in any column, provided that if one such additional suppression is chosen in a column containing no suppressions, then at least one other such additional suppression is chosen in this column also.

To illustrate the number of such "minimal" patterns that can be formed when $R = C$ and the original suppressions are each in a different row and column of the table, the number of "minimal" protecting patterns equals the Rth Menage number (also known as the number of *derangements*), which, for $R \geq 6$, is approximately $R!/e$, where $e = 2.71828 \ldots$ is the base of the system of natural logarithms. Among these many minimal patterns an adequate pattern usually exists. Indeed, this is often true even if the search is restricted to candidate complementary cells Y satisfying Theorem 2. A future direction for solving the complementary suppression and validation problems operationally also exists within the framework of mathematical programming.

6. CONCLUDING COMMENT

The preceding pages illustrate the mathematical and computing complexity of the cell suppression problem. Although of legitimate intellectual interest, in any practical application these problems must be examined in terms of their significance and consequences. For example, in applications in which respondents do not contribute to more than one cell in a cell union, or do so rarely, choice of sensitivity criteria that reflect this situation will spare considerable effort that might otherwise be necessary.

This article attempts to highlight the interrelationships between the processes of disclosure definitions, subproblem construction, complementary cell suppression, and validation of the results. In any particular application, integration of these concepts, such as within mathematical programming framework, is necessary to attain proper balance between disclosure risk and the society's data needs and between current methodology and available resources.

[*Received January 1978. Revised October 1979.*]

REFERENCES

Cox, Lawrence (1975), "Disclosure Analysis and Cell Suppression," *Proceedings of the American Statistical Association*, Social Statistics Section, 380–382.

——— (1977), "Suppression Methodology in Statistical Disclosure Analysis," *Proceedings of the American Statistical Association*, Social Statistics Section, 750–755.

——— (1978a), "(−1, 0, +1)-Linear Estimates of Suppressed Entries in a Simple Table," U.S. Bureau of the Census, unpublished draft.

——— (1978b), "Automated Statistical Disclosure Control," *Proceedings of the American Statistical Association*, Survey Research Methods Section, 177–182.

——— (1979), "Linear Sensitivity Measures in Statistical Disclosure Control," *Proceedings of the American Statistical Association*, Social Statistics Section, 634–639.

Cox, Lawrence, and Sande, Gordon (1979), "Techniques for Preserving Statistical Confidentiality," *Proceedings of the 42nd Session of the International Statistical Institute*, in press.

Dantzig, George (1963), *Linear Programming and Extensions*, Princeton, N.J.: Princeton University Press.

Kuhn, Harold, and Tucker, Albert (eds.) (1956), *Linear Inequalities and Related Systems*, Princeton, N.J.: Princeton University Press.

Sande, Gordon (1977), "Towards Automated Disclosure Analysis for Enterprise Based Statistics," Statistics Canada, unpublished draft.

U.S., Department of Commerce (1978), "Statistical Policy Working Paper 2: Report on Statistical Disclosure and Disclosure-Avoidance Techniques," Washington, D.C.: U.S. Government Printing Office.