Session 1.

Ensuring Data Confidentiality

A DISCLOSURE LIMITATION METHOD FOR TABULAR DATA THAT PRESERVES DATA ACCURACY AND EASE-OF-USE

Lawrence H. Cox
National Center for Health Statistics
Centers for Disease Control and Prevention
LCOX@CDC.GOV

Ramesh A. Dandekar
Energy Information Administration
Department of Energy
Ramesh.Dandekar@EIA.DOE.GOV

Abstract

Disclosure limitation in tabular data traditionally has been accomplished by subjecting cell values to any of three methods: rounding, perturbation, or complementary cell suppression. If outputs are two-dimensional tables arranged independently or hierarchically, all three methods rest on sound theory and efficient computational algorithms that can be implemented up to the level of a census or large survey. Beyond two-dimensions, the reverse is true: the close connection between mathematical theory and efficient computational algorithms breaks down and computational requirements escalate. Each method is effective for disclosure limitation in contingency (count) data wherein disclosure is associated with small cell values. For magnitude data such as sales or expenditures data, disclosure can be associated with cell values of any size, rendering rounding and perturbation ineffective or inferior to cell suppression in most situations. Unfortunately, cell suppression can create patterns of missing cell data that may destroy information important to certain users and are difficult to analyze properly by all but sophisticated users. These factors create a complicated and undesirable situation from both a statistical and policy perspective: none of the current methods assures the creation of a complete, accurate, disclosure-limited data product that is as easy to use as the original (pre-disclosure limited) data, created in a flexible manner at reasonable computational effort. We present an alternative method designed to preserve these properties. We refer to this method *controlled adjustment of tabular data*, or controlled tabular adjustment. It is a method for large-scale controlled data perturbation based on linear programming. We discuss issues of expected importance to data producers and data users and illustrate how these can be accommodated flexibly within the controlled tabular adjustment framework.

1. Introduction

In this paper, we outline a new methodology for disclosure limitation in statistical data presented in tabular form. We focus on properties and applicability of the method and omit technical details, available in Dandekar and Cox (2002). Similarly, we do not repeat definitions or review the extensive literature on statistical disclosure limitation, also available in Dandekar and Cox (2002) and elsewhere, except as needed to provide relevant context in which to examine the new method. Precise development of terminology and concepts is eschewed to favor a descriptive presentation.

A Typical Situation

A National Statistical Office (*NSO*) collects data on individual entities (persons, businesses, farms, hospitals, ...), processes the data, and releases information in the form of *statistical data products* to the public and decision makers. Traditional data products are large systems of predetermined tabulations (as from the Economic Censuses), public use or specialized microdata files (as from demographic surveys), and special tabulations. Emerging new forms of data release include tabular or analytical (e.g., regression) output from statistical data base query systems.

Statistical disclosure occurs if a third party (the *intruder*) can use released data products to associate an individual entity with either:

- a tabulation cell (in tabular data from a census or survey)
- an individual record (on a microdata file)
- a response to a query (to a statistical data base query system)

and, - can deduce or infer one or more of the entity's confidential attributes.

This has been called *attribute disclosure*. In certain situations or programs (e.g., Statistics of Income), association alone may constitute disclosure (*identity disclosure*). The NSO usually takes into account the possibility that the intruder will use *auxiliary information* (public knowledge, matching file, ...) to achieve disclosure, but often must do so without complete knowledge of sources or specifics pertaining to this (potential) information. An exception is tabular economic statistics wherein the best informed intruder is often a competitor contributing to one or more tabulations involving the *target* of the disclosure.

Achieving Disclosure

Confidential attributes are often deduced via mathematical manipulation of released data. Tabular data are organized by categorizing respondent data within *elementary tabulation cells* defined by one or more variables (e.g., Age by Race by Sex in the Current Population Survey, North American Industry Classification System (NAISC) by Metropolitan Statistical Area (MSA) in the Economic Censuses, Age by Sex by International Classification of Disease (ICD) code in national health surveys). Each elementary tabulation cell is assigned a *cell value* corresponding to a statistic of interest. For *categorical data*, the cell value equals the number of respondents in the cell. For *magnitude data*, the cell value equals the sum over all respondents in the cell of a quantity of interest (e.g., income, number of doctor visits, total quantity of a commodity shipped by a manufacturer). Cell values of elementary tabulation cells are then aggregated to produce values for tabulation cells at successively less refined levels of detail (e.g., for States, the entire United States, larger industry groupings, broader Age categories). For survey (as opposed to census) data, there may be an intermediate step at which the individual data are weighted. Because this organization–from individual data to elementary tabulation cells to more general tabulation cells–is based on addition, it can be realized mathematically as a system of linear equations.

Disclosure occurs if the intruder can work backwards from aggregated data to deduce individual respondent data. In certain cases, this can be accomplished by linear algebra. By the same token, disclosure occurs if the intruder can estimate individual respondent data to within an unacceptable narrow (*prohibited*) range (what is meant by "narrow" is determined by the statistical agency and often varies from agency to agency and sometimes from survey to survey). Narrow estimation, whenever possible, can be accomplished by linear programming.

Disclosures as above are achieved by deterministic means, so that respondent data are identified within a range. *Probabilistic disclosure* determines if, within an acceptable range, there is high probability that respondent data lie within a smaller prohibited range. Probabilistic disclosure is only beginning to be addressed in the literature and is beyond the scope of this paper.

The paper is organized as follows. Section 2 describes typical mechanisms for quantifying statistical disclosure in tabular data products. The new method is introduced in Section 3. Two questions are central in the evaluation of a disclosure limitation method. Does the method provide the required degree of disclosure limitation, that is, has it reduced the risk of disclosure to a sufficient extent? This question must first be answered in the affirmative. The second question is then: Has the method preserved important analytical properties of the data? The first question is addressed in Section 3, the second in Section 4. Section 5 provides concluding comments.

2. Quantifying and Limiting Statistical Disclosure in Tabular Data

Quantifying Disclosure

For categorical (count) data, statistical disclosure occurs when an individual can be correctly associated with a specific set of characteristics or *attributes*. The concern is that known or publicly available attributes of the respondent (e.g., sex, age category, profession, industrial classification, geographic area where a person lives or business, medical or insurance services are offered) can be used to identify the respondent in the data and from there link the subject to its confidential attributes (e.g., illegal drug use, income category, disease incidence, corporate cost, sales or employment practices information, medical insurance costing or reimbursement policies). A clear problem exists if the respondent is categorized in a tabulation cell containing only a small number of respondents, viz., the cell value is *small*. Or, further, if a small cell or cell complement can be so-identified. What is meant by "small" is determined by the policies and practices of the NSO and/or survey. For example, the U.S. Census Bureau has in the past used values such as "5" for census data and "15" for survey data. Statistics New Zealand and the Statistics of Income Program use "3".

Disclosure in categorical data is thus defined by a *threshold rule*: a cell or cell combination (or complement) is a disclosure if its value is less than a predetermined threshold value n (e.g., n = 3, 5, 15). Consequently, narrow estimation is defined to be an estimate of a cell value, computed by deterministic means such as linear programming, that is less than n. Because cells for which there are no respondents or data, and consequently have cell value equal to zero, are often well-known, *zero cells* are typically exempted from the notion of "small". Typically, the NSO makes the numeric value of n publicly available.

This rule can be expressed quantitatively in the following manner. A nonzero cell or cell combination **X** is a *primary disclosure cell* under the *n-threshold rule* if:

$$S(X) = n - m > 0$$

where m denotes the number of respondents in the tabulation cell or cell combination. The prohibited range under the n-threshold rule is thus the interval (0, n). Because inferences equal to zero or n are permitted, this is an *open interval*, viz., the endpoints are excluded. Conversely, a range estimate for a cell that strictly contains the prohibited interval must be acceptable.

For magnitude data, disclosure amounts to narrow estimation of a quantitative attribute corresponding to the respondent. For, e.g., manufacturing or business data, it is often easy to associate individual respondents to particular tabulation cells (e.g., type of good manufactured or goods sold and location of factories or retail outlets are well known). The NSO may consider this information to be publicly available. What the NSO must protect from disclosure are the quantitative attributes of the respondent (e.g., sales, cost or pricing data). Here disclosure is a bit more complex because the most likely intruder may be a competitor whose data are also contained in the cell total. It is instructive to proceed from an example from manufacturing statistics.

Assume that four companies contribute their individual Total Value of Shipments (TVS) to the Manufacturing Census, and that the respective contributions, measured in some appropriate units, are 55, 40, 3 and 2 units. The *true* cell value is therefore 55+40+3+2 = 100 units. If the cell value is published, Contributor #2 can subtract its contribution (40) from the published total (100) to infer that its largest competitor had TVS at most 60 units. This estimate is therefore accurate to within 9% of the actual contribution. If the NSO regards 9% as "too close" (and, typically, an NSO would do so), then releasing this cell value would result in disclosure (to Contributor #1 by Contributor #2).

A typical disclosure rule for magnitude data is the *p-percent rule*, illustrated above: no estimate of any respondent by another respondent can come within p-percent of the first respondent's contribution to the cell. In contrast to categorical data where the threshold *n* is made publicly known, the NSO typically keeps the value *p* confidential as an additional safeguard to confidentiality.

It results that the greatest threat to a respondent by another respondent or third party is that illustrated above: where Contributor #1 is the target and Contributor #2 is the intruder. The p-percent rule can be represented quantitatively in the following manner. A cell **X** is a disclosure under the p-percent rule if:

$$S(X) = \sum_{i=3}^{m} x_i - (p/100)x_1 > 0$$

where x_i denotes the contribution of the i-th largest respondent (ordered from largest to smallest) to cell **X**. For simplicity, we assume all respondent contributions are nonnegative. Clearly, all cells with only one or two respondents satisfy the rule.

The prohibited range for primary disclosure cell **X** under the p-percent rule follows directly from the quantitative disclosure rule, as follows. The upper endpoint of the prohibited range should be the smallest value of a (hypothetical) cell containing **X** for which the quantitative rule fails to hold. This value is precisely the cell value of **X** plus S(X). Computation of the lower endpoint of the prohibited range is more complicated, and NSOs often replace it by the cell value of **X** minus S(X).

Limiting Disclosure

There are several disclosure limitation methods available for tabular data. For convenience, we characterize these either as *perturbative* methods or *suppression* methods.

Perturbative methods modify some or all of the true cell values to make it impossible or unlikely that the intruder can narrowly estimate the original primary disclosures. *Random perturbation*, which has been practiced by NSOs in the United Kingdom, amounts to adding or subtracting a small randomly determined integer value (possibly zero) to original cell values. In this way, the intruder cannot with certainty conclude that a published small value corresponds to a true small value. The NSO may or may not make the perturbation values and/or the perturbation probabilities publicly known.

Rounding is a form of perturbation for which all cell values are rounded either down or up to an adjacent multiple of some rounding base B (under the n-threshold rule, B = n). In this way, the intruder cannot with certainty conclude that a published cell value corresponds to a small original value. As B = n and as it is obvious when data have been rounded, no attempt is made to conceal the rounding base B. Random rounding is performed using a randomization method that ensures that expected values of rounded entries equal original entries. The rounding probabilities are uniquely determined, so no attempt is made to conceal them. A variant is *minimum distance rounding*, e.g., with respect to minimum sum of squared differences between rounded and original entries.

Simple conventional rounding (e.g., base B=5, round 0, 1, 2 down to 0 and round 3, 4, 5 up to 5) does not preserve additivity (e.g., $3 + 4 = 7$ but $5 + 5 \neq 5$). For one- and two-dimensional tables, random perturbation and random and minimum distance rounding can be performed in a manner that preserves additivity. This is *controlled rounding* (Cox 1987). Unfortunately, controlled rounding is not always possible in three- or higher-dimensions or for linked tables.

Complementary cell suppression is a third disclosure limitation method for tabular data. Under complementary suppression, primary disclosure cells are suppressed from publication, viz., the corresponding values are replaced by a suppression symbol, denoted **D**. Because (narrow) estimates of suppressed cell values can be obtained by manipulating aggregation equations between cell values, it is often the case that additional, nondisclosure cells, called *complementary suppressions*, must also be suppressed to prevent narrow estimation of primary disclosures. Combining two or more data categories (known as *collapsing*) can be viewed as (wholesale) complementary suppression. Complementary suppression is a complex theoretical, computational and operational undertaking.

Perturbation, rounding and suppression all are suitable disclosure limitation methods for categorical data. Because perturbation and rounding produce more useable results, these methods generally are preferred to suppression for for disclosure limitation in contingency tables. As an illustration, Figure 1 presents an original contingency table under the 5-threshold rule, alongside the table after controlled rounding and complementary cell suppression.

Perturbation and rounding in general are ineffective for disclosure limitation in magnitude data, for two reasons. First, magnitude data typically are skewed, necessitating changes of different magnitudes to individual cells. Second, perturbation and rounding are designed to introduce small changes into cell values, whereas rules like the p-percent rule often dictate larger changes (e.g., 5%-20% of cell value). Consequently, complementary cell suppression has become a defacto standard for

disclosure limitation in tabular magnitude data, despite it being difficult to perform and control, its computational demands, and its removal of useful data and thwarting statistical analysis. It is not that data producers or users like complementary suppression--there simply has been no realistic alternative.

20	11	28	2	19	80
12	12	21	3	12	60
39	11	3	20	17	90
4	1	13	20	2	40
75	35	65	45	50	270

20	10	30	0	20	80
15	10	20	5	10	60
40	10	5	20	15	90
0	5	10	20	5	40
75	35	65	45	50	270

20	***D***	28	***D***	19	80
D	12	21	***D***	12	60
39	11	***D***	20	***D***	90
D	***D***	***D***	20	***D***	40
75	35	65	45	50	270

Figure 1: *Original, Rounded, and Suppressed Two-Dimensional Contingency Table*

Figure 2 illustrates complementary cell suppression. Assume that the six cells in bold are primary disclosures. To simplify understanding, assume each primary disclosure requires protection equal to 10% of its value, viz., the prohibited range for a cell of value 200 is the open interval (180, 220). Alongside the original table is one possible suppression pattern to protect this table. In lieu of suppression symbols **D**, we provide best-possible (*exact*) interval estimates for suppressed cells. Note that, for the six primary disclosure cells, each exact interval contains the prohibited range, as required.

200	40	50	**200**	120	610
20	70	60	**100**	120	370
40	90	**250**	100	30	510
100	**150**	30	80	**150**	510
360	350	390	480	420	2000

200	***[0,60]***	50	**[180,240]**	120	610
[0,60]	70	60	**[60,120]**	120	370
40	90	**[100,280]**	100	***[0,180]***	510
60,120]	**[130,190]**	***[0,180]***	80	**[0,180**	510
360	350	390	480	420	2000

Figure 2: *Table of Magnitude Data Before and After Complementary Cell Suppression*

Complementary cell suppression leaves some data fixed but removes other data. For the naive user, the missing data appear to be removed entirely. The more sophisticated user could compute exact interval estimates for the missing data (see Figure 2) and impute the missing values based on these intervals. Indeed, some practitioners, e.g., Gordon Sande, have suggested that NSOs release the exact intervals as in Example 2 to assist all users. Sophisticated users might employ missing data strategies, e.g., the E-M algorithm, to impute the missing data. Indeed, a largely unexplored problem with cell suppression is the ability of such strategies to narrowly estimate original (confidential) values.

3. The New Method–Controlled Tabular Adjustment

Our objective is to develop a method for statistical disclosure limitation in magnitude data that preserves analytical properties of original data and offers acceptable theoretical and computational properties and performance in multi-dimensional settings. It should be an improved alternative to complementary cell suppression. An useful analogy is between controlled rounding and cell suppression in two-dimensional contingency tables. Controlled rounding can be performed optimally and efficiently in two-dimensions and produces a table "nearby" the original table devoid of missing entries. Suppression is more difficult to perform optimally and, while keeping some values fixed, removes other values. Most would agree that, for two-dimensional contingency tables, controlled rounding is an improved alterative to complementary suppression.

Our objective is to provide analogous improvement for magnitude data in two and higher dimensions. Applications can be as large as a national census or survey such as Censuses of Manufacturing or Retail Trade that contain many thousands of tabulation cells, at many levels of aggregation (viz., totals/subtotals/sub-subtotals/...../detail), and span several to many logical dimensions (viz., classification variables such as geography, NAICS, size categories, ...). Relying on heuristic methods, complementary cell suppression has been made to work in such applications since the 1970s at the U.S. Bureau of the Census and Statistics Canada but at the cost of *oversuppression* of data and patterns of missing data that can be difficult to analyze.

From the outset, it should be clear that our proposed method is NOT complementary cell suppression (**CCS**). Both methods are designed to provide disclosure limitation in tabular data. As we present our method as an improved alternative to complementary cell suppression, it is worthwhile to summarize the principal features of CCS. We focus primarily on magnitude data, that being the area most in need of an alternative to suppression.

Disclosure in tabular data is based on the risk of identifying confidential information pertaining to an individual respondent. Disclosure rules characterize this risk by labeling each tabulation cell either as a *primary disclosure cell* or not. Using the disclosure rule, each tabulation cell **X** considered for release is examined for disclosure. For categorical data, the disclosure rule might be the n-threshold rule, e.g., n = 5. For magnitude data, the disclosure rule might be the *p-percent rule*.

To characterize the disclosure risk associated with publishing primary disclosure cells, a *protection interval* $[L_X, U_X]$ is assigned to each primary cell **X**. The protection interval is computed directly from the disclosure rule and the contributor data corresponding to cell **X**. Estimates of the value of **X** breaching this interval are *disclosures*; interval estimates of the value of **X** that contain (are at least as broad as) the protection interval are *acceptable*. This characterization is important–it provides both quantification of risk and a decision rule for determining when *sufficient disclosure limitation* has been achieved. Complementary cell suppression then can be performed to achieve sufficient disclosure limitation. A simplified synopsis of complementary suppression is as follows.

Under complementary suppression, each primary disclosure cell is suppressed from publication (and replaced by a symbol **D**). The system of tabulation equations naturally defines a system of linear equations **S** among the cell values, in which the value of a cell **X** corresponds to a variable **x**. Initially, variables corresponding to the non-primary disclosure cells are replaced by their true values, so that only the primary disclosure cell values are represented by variables. Linear programming analysis can be applied to the system **S** to obtain exact interval estimates $[\min_S \{x\}, \max_S \{x\}]$ of the value of each suppressed primary disclosure cell **X**. If any of these intervals fails to contain the corresponding protection interval, then disclosure occurs. It is then necessary to suppress additional, *nondisclosure cells* until all protection intervals are contained in the corresponding exact intervals. This amounts to replacing selected true values of non-primary disclosure cells with variables until the exact interval test is met for each primary cell. We do not describe this process further, except to emphasize that it is equivalent to solving a typically large integer linear program and that the computational effort and time required to do so can be prohibitive. From the standpoint of analysis, once complementary suppression is complete, most users can only guess values of primary disclosure cells at best to within the protection limits, and, for nondisclosure cells, to within arbitrary limits.

Returning to Figure 1, after attempting complementary cell suppression in the rightmost table, exact interval estimates are given by Figure 3. Note that two of these estimates (both equal to [0, 4]) actually fail the exact interval test (because their right-hand endpoints lie in the protection interval), necessitating further disclosure analysis and complementary suppression (not shown here).

20	[8,12]	28	[1,5]	19	80
[11,15]	12	21	**[0,4]**	12	60
39	11	[1,16]	20	[4,19]	90
[1,5]	**[0,4]**	[0,15]	20	[0,15]	40
75	35	65	45	50	270

Figure 3: *Exact Interval Estimates After Complementary Cell Suppression*

We next describe the new disclosure limitation method for magnitude data, using the example provided in Figure 4. Assume that the cells in Figure 4 represented in **boldface** are the primary disclosure cells and, for ease of understanding, that the protection interval corresponding to each primary disclosure cell is the interval corresponding to +/- 10% of the true cell value **x**, viz., the interval (0.9x, 1.1x). The endpoints of a protection interval are called the lower/upper *protection limits*.

200	40	50	**200**	120	610
20	70	60	**100**	120	370
40	90	**250**	100	30	510
100	**150**	30	80	**150**	510
360	350	390	480	420	2000

Figure 4: *Table of Magnitude Data with Six Primary Disclosures (Protection Required for Each Primary Disclosure = +/- 10%* of Cell Value)

The new method is based on adjusting many and potentially all cell values in a manner that: 1) provides sufficient disclosure protection for the primary disclosure cells, 2) preserves the additive structure of the tabulations, and 3) minimizes individual adjustments and any of several sensible measures of overall adjustment towards preserving analytical properties of the data. This can be accomplished in many ways that are explored in the next section. As a starting point for introducing the new method, here we offer the following adjustment schema:

- replace the value of each primary disclosure cell with a *safe value*, viz., a value that does not represent disclosure (this is the *instantiation* step); an obvious choice is
  * a value at or beyond either of the primary cell's lower or upper protection limit
- assign nonnegative variables y^-_i, y^+_i to each non-primary cell value or total **i**
  * these variables represents potential downward/upward adjustment to the cell value
- represent the additive tabulation relationships (viz., from detail to sub-totals, sub-totals to higher-level sub-totals, ..., and ultimately to grand total) as a *system of linear equations*, denoted **S**
- augment **S** with *capacity constraints* on the cell adjustment variables **y** to ensure that values of nondisclosure cells do not change too much; sensible capacity constraints
  * constrain each **y** to be within a small percentage of the true cell value
  * constrain each **y** to be within estimated measurement error of the true cell value
- impose a linear *cost function* **c** on **S** that represents a sensible measure of overall change to the data; standard possibilities include
  * sum of absolute deviations from original values
  * average percent deviation from original values
  * sum of logarithms of 1 + deviations
- use linear programming on **S**, **c** to instantiate remaining values in a manner that
  * assures all additive tabulation relationships are preserved
  * minimizes the measure of overall change **c**

The linear program performs these tasks automatically. Linear programs are computationally efficient even for large problems. Massively large problems require specialized techniques.

The schema outlines a method for *controlled tabular adjustment* (**CTA**). CTA transforms a tabular system *with* disclosures to one *without* disclosures. The schema describes the method sufficiently for understanding the remainder of this paper. A formal mathematical statement of the CTA schema follows. Understanding this model is not required to follow the remainder of the paper.

Mathematical Model for Optimal Controlled Tabular Adjustment

Notation

$i = 1, \ldots, p$: denote the p primary disclosure cells
$i = p+1, \ldots, n$: denote the n-p nondisclosure cells
M = coefficient matrix of the tabular system **S**

I_i = binary (zero/one) variable denotes selection of lower/upper protection limit at which to instantiate primary disclosure cell $i = 1, \ldots, p$
y_i^- = potential downward adjustment to cell value i
y_i^+ = potential upward adjustment to cell value i

$LPROTECT_i$, $UPROTECT_i$ = lower/upper deviation required to protect primary disclosure cell $i = 1, \ldots, p$
* these values are derived directly from the disclosure rule and the cell contributions
LB_i, UB_i = lower/upper bound (capacity) on downward/upward change to cell $i = 1, .., n$
* these values are determined by analytical or data quality requirements
c_i = cost per unit change in cell i
* these values are determined by NSO policy/practice

Mixed Integer Linear Program (MILP) for CTA (simplified)

Minimize: $\sum_i c_i(y_i^- + y_i^+)$
Subject to:

For $i = 1, \ldots, n$:

$$\mathbf{M}(\mathbf{y}^+ - \mathbf{y}^-) = \mathbf{0}$$
$$0 \le y_i^- \le LB_i$$
$$0 \le y_i^+ \le UB_i$$

For $i = 1, \ldots, p$:

$$y_i^- = LPROTECT_i * (1 - I_i)$$
$$y_i^+ = UPROTECT_i * I_i$$
$$y_i^-, y_i^+ \ge 0, \; I_i \in \{0, 1\}$$

Figure 5 illustrates a possible controlled tabular adjustment of the table with disclosure presented in Figure 4. This solution was obtained "by-hand" and therefore is not optimal. Using the cost function equal to absolute-sum-of-deviations, viz., $c(y) = \sum_i (y^-_i + y^+_i)$, an optimal CTA is given in Figure 6.

200	40	50	**200**	120	610
20	70	60	**100**	120	370
40	90	**250**	100	30	510
100	**150**	30	80	**150**	510
360	350	390	480	420	2000

195	35	55	**220**	115	620
30	65	65	**90**	125	375
45	95	**225**	105	35	505
90	**165**	35	75	**135**	500
360	360	380	490	410	2000

Figure 5: *Table of Magnitude Data with Six Primary Disclosures, Before and After CTA*

189	36	45	**220**	120	610
22	70	56	**90**	132	370
37	81	**275**	90	27	510
110	**165**	27	73	**135**	510
358	352	403	473	414	2000

Figure 6: *Optimal Controlled Tabular Adjustment of Figure 4 With Respect to Minimum Sum-of-Absolute-Deviations*

The sum-of-absolute deviations in Figure 5 equals 240; the optimal value, from Figure 6, equals 198. For simplicity, no capacity constraints were imposed. There are many adjustments with this optimal cost. A different cost function can produce a different optimal solution. In the next section we argue that, for practical purposes, there is little discernible difference between two adjustments like those in Figures 5 and 6.

The mathematical model describes a *mixed integer linear program* (**MILP**) because the variables I_i are binary integers. Integer programs are very hard to solve efficiently, except for small problems. In general, we do not recommend the pursuit of an optimal MILP solution. Instead, the use of *heuristic* methods

to instantiate the primary disclosure cell values is recommended. This reduces the problem to linear programming. Heuristics are discussed in the next section and in detail in Dandekar and Cox (2002). Comparisons with optimal solutions are made in Cox and Kelly (2003).

In summary, controlled tabular adjustment, produces a system of tabular cell values that

- is additive to all sub-totals and totals
- for nondisclosure cells, the instantiated values
  * are close to original values individually
  * minimize an overall measure of deviation from true values
- for primary disclosure cells, the instantiated values
  * do not represent disclosure
  * are better than what the user gets under **CCS**
- is as easy to analyze as original data

This new disclosure limitation methodology

- is computational efficient
- can be repeated many times using different constraints and costs to simulate/examine a range of releasable data tabular products
- consequently, can be run, examined, and fine-tuned to specific survey conditions by NSO subject-matter analysts
- obviates the need for complementary cell suppression

Whereas complementary cell suppression is a *turn-key system* in that it allows little interaction by subject analyst, controlled tabular adjustment is more of an *expert system* or *expert assistant* (such as in medical diagnosis or architectural design) to augment the capabilities of the subject analyst. In the next section we examine some of the potential pros and cons of this new method and its potential for preserving analytical properties of the original data.

4. Properties of the CTA Method and Data Analysis Issues

This discussion is organized around questions that naturally arise.

Each disclosure primary cell is instantiated with a value at or near its lower or upper protection limit. Is this easy to do? Does how this is done make any difference?

As discussed in Section 3, instantiating the primary cells optimally requires solving a mixed integer linear program. This is computationally demanding for small problems and impossible for large problems. The use of heuristics for the instantiation is indicated. Random instantiation of the primary cells can be done quite easily. Unfortunately, experience (Cox and Kelly 2003) demonstrates that random solutions tend not to be close to optimal. However, computing, say 100 randomly instantiated solutions and choosing the best one often works well.

Other heuristic approaches include ordering the primary cells from largest to smallest value and assigning the lower/upper deviation in alternating fashion. More are emerging.

It is important to note that the meaning of optimality in this context is less clear than for example for mathematical optimization problems based on an actual dollar cost. Consider Figures 5 and 6. Is there really a meaningful difference between the two adjustments? In the literature and among practitioners, there is no consensus on the form of "best" cost function would take (e.g., minimize total absolute deviations, or minimize total percent deviation). Whereas an optimal solution establishes a gold standard mathematically, it cannot incorporate all the subjective information an analyst might incorporate. We expect that the ability to produce a variety of near-optimal solutions for analyst review and refinement will be seen as more valuable than exhibiting a mathematical optimum.

Primary disclosure cells may be changed quite a bit. Won't this bias data analysis?

Certainly changes other than small changes to a cell value biases that value and enough changes of this magnitude can bias analysis of a subdomain or the entire data set. Changes to primary disclosure cell values are determined by the disclosure rule and the cell data, and percent deviation will vary from cell to cell and survey to survey. Under typical NSO scenarios, the percent deviations are likely to be in 0% to 15-20% range. Changes at the upper end of this range certainly are liable to create bias. Empirical studies have shown that, without further attention to this issue, a small bias is introduced in the regression of instantiated values on original values. A worst case is would be if every primary disclosure value were adjusted upwards by a fixed percentage p, for then the regression coefficient would equal $(1 + p/100)$. But, under this scenario, correlation would equal one. Empirical studies demonstrate small change in correlation among instantiated and original primary disclosure values.

As the only alternative to CTA for disclosure limitation is complementary cell suppression, it is appropriate to assess the effects on analysis of CTA in comparison to those of complementary suppression. Complementary cell suppression forces the user to estimate the true value only within an interval at least as broad as the protection interval. If the user could estimate any closer value with confidence, then confidentiality would be breached. Therefore, instantiation of either the lower or the upper protection limit for each primary cell leaves the user with no more bias than suppression. Indeed, CTA provides the user with a unique value, enabling analysis by even the most unsophisticated user.

Still, this could result in bias. Closer examination reveals that the NSO can in fact release a closer value that still is safe. The user (and the intruder) have no way of knowing whether the original value was instantiated down or up from the true value. Thus, releasing a value in the protection interval provides the intruder no reliable means to obtain a narrow interval estimate the contribution of a target respondent. (An exception is single-respondent cells that must be treated separately.) The NSO could instantiate values for primary disclosure cells by random selection from values in the protection interval with respect to an appropriate distribution. This can be done with little or no bias. Because this results in smaller adjustments to primary disclosure cells, it requires smaller changes to individual cells and overall, thereby better preserving analytical properties of the data set. This approach does raise a policy issue as the perception that the NSO is releasing nearby values may be problematic.

Can CTA assure only small changes to nondisclosure cells?

The NSO can constrain changes to nondisclosure cells to be as small as desired. If solutions satisfying these requirements exist, the linear program will find them. If solutions do not exist, because this method is computationally efficient, it is then possible to either reinstantiate the primary cells and run the linear program again, relax some or all of the variable constraints and run again, or both.

It is important to note that constraints can be variable-specific, meaning that a variable for which no change is appropriate can be fixed at its original value and/or looser constraints can be assigned to unimportant/unreliable cell values.

In the typical case where the disclosure cells do not dominate the system, tightly constrained solutions should be available. A strong advantage of our approach is that all of these considerations can be expressed formally within a single linear program that in many situations can be run multiple times to represent different scenarios or desiderata.

Consider also the obverse: if it is inordinately difficult to balance protection with efficient selection of local changes in CTA, then it must be at least this difficult to obtain a pattern of complementary suppressions that is useful/tractable for analysis.

What are the likely effects of CTA on data analysis?

It is important here to acknowledge that first one must specify "which analysis": analytical scenarios and issues tend to be data-dependent. A census or survey offers myriad possibilities for analysis. Census and survey data are also subject to various sources and levels of error, whose effects on analysis are largely unknown. An approach as we have offered that minimizes or controls change at both the individual cell and overall is an important feature.

Change must be examined at three levels: for the primary disclosure cells, for the nondisclosure cells, and overall. Effects on the primary disclosure cells were discussed in an earlier subsection. These effects are no worse than for complementary suppression and, if our suggestions are followed, can be improved considerably by judicious choice of instantiations.

Nondisclosure cells are changed by only a small percentage. Empirical studies show that regressions and correlations are good. Arguably, if changes to nondisclosure cells are confined to within measurement error, then original and adjusted data are for all intents and purposes indistinguishable statistically.

In most settings, the primary disclosures are only a small part of the overall tabulations, and do not tend to dominate the larger values. This results in very small change to regressions and correlations among all cells, borne out by empirical studies.

CTA provides complete data, so analysis is as easy and simple as for original data. The ability to control change to individual cells allows analytically unique or important cells to be treated favorably. Conversely, less important cells can be allows to vary to a greater extent. If our suggestions are followed, changes to primary disclosure cells are no worse than complementary suppression, easier to deal with analytically, and may be expressed as random draws from known distributions.

The release of model-generated microdata in lieu of original data for disclosure limitation purposes has been suggested. How does this methodology relate to that?

The difficult thing to control in tabular data is the tabulation structure among the cells and cell values. Models for *synthetic microdata* based on microdata, as suggested by Rubin, do not have to contend with these issues at their typical levels of complexity. It has been suggested that synthetic microdata could be released under the *multiple imputation* paradigm by releasing multiple versions of the tabular system. For tabular data, this is likely to reverse disclosure limitation as the tightly defined cell and tabulation structure would force averages across multiple files of "synthetic tabulations" to be very near original cell values, thereby increasing risk of disclosure.

Are there other potential approaches for controlled tabular adjustment?

Doubtless other approaches will emerge. One statistical approach would be to develop algorithms for *iterative proportional fitting* in complex tabular settings. A potential drawback is that limited empirical experience indicates that predicted values tend to be closer to true values. Also, development of such algorithms for complex, multi-dimensional tabular systems may be tricky.

Linear programming, as used here, finds *extremal solutions* among all possible (*feasible*) solutions. Except for purposes of optimizing the linear cost function, there is no particular reason to favor extremal solutions. Indeed, although very efficient, there are limitations to linear programming vis a vis problems size. An approach that sought or exploited feasible solutions in general would be advantageous in these situations. Kelly et al.(2003) are exploring search algorithms for moving from feasible to better or near-optimal solutions using *Tabu search*. Direct search procedure have the additional advantage of lifting the requirement that cost functions be linear. This enables comparison of original and adjusted data based, e.g., on correlation, chi-square, etc.

5. Concluding Comments

Controlled tabular adjustment potentially offers an improved alternative to complementary cell suppression in terms of data analysis, simplicity of the theoretical model, interaction of the methodology with subject matter analysts, flexibility in use and operational/computational performance. Instantiation of values for multiple-respondent primary disclosure cells from known distributions and of nondisclosure values within measurement error would assure both confidentiality protection and consistency of analytical results.

We have offered a new approach to disclosure limitation in tabular data that enables variations and refinements to meet a wide range of survey, analytical and computational settings. It is the first step in replacing data suppressed to preserve confidentiality in tabular magnitude data released by NSOs with nonconfidential data suitable for analysis. Future research and evaluation areas for this new methodology include acceptance of synthetic data products by producers and users, good heuristics to obtain near-optimal solutions, integerization of continuous outputs for contingency tables, examination of effects on data analysis, limitations/opportunities for interaction with subject analysts, identification/development of supplementary information to improve analytical outcomes and account for bias, exploring limitations of/alternatives to linear programming solutions (e.g., nonextremal feasible solutions), and incorporation of nonlinear cost functions related to statistical analysis.

Our approach utilizes linear programming as a means to preserve additive tabular structure. Analytical properties are preserved and biased controlled to the extent possible by imposing appropriate constraints on individual cell adjustments. Optimality of the final solution in many cases is only an artifact in the sense that no meaningful difference can be discerned between optimal and near-optimal solutions, including nonextremal feasible solutions. This flexibility enables the development of other methodologies, including branch-and-bound and direct search, to perform controlled tabular adjustment. We look forward to further developments and refinements for controlled tabular adjustment.

Disclaimer

The opinions expressed herein are solely those of the authors and should not be interpreted as representing the policies or practices of the Centers for Disease Control and Prevention, the Energy Information Administration, or any other organization.

References

Cox, LH (1987). A constructive procedure for unbiased controlled rounding. *Journal of the American Statistical Association* **82**, 520-524.

Cox, LH and J Kelly (2003). An empirical study of heuristic and optimal methods for controlled tabular adjustment. Manuscript.

Dandekar, RA and LH Cox (2002). Controlled tabular adjustment: an alternative to complementary cell suppression for disclosure limitation of tabular data. Manuscript.

Kelly, J et al. (2003). Controlled tabular adjustment using tabu search. Manuscript.

Issues and Impediments to Expanding Access to Confidential Statistical Agency Data: Restricted Data and Restricted Access

Stephen H. Cohen
Bureau of Labor Statistics

Wilbur Hadden
National Center for Health Statistics

Abstract

The Federal Statistical Agencies collect a wealth of confidential economic, demographic and social data. These data are collected to meet requirements in legislation or the code of federal regulations. The agencies publish estimates from that data in various tabular forms on paper or on the Internet. However, analysts still are interested in the wealth of potential additional tabulations that are not published by agencies and in developing statistical models of the data.

Historically, responding to these interests the agencies publish micro data sets for demographic surveys, but the agencies are limited in what data can be released by requirements to protect the confidentiality of data providers and survey respondents.

More recently, agencies have created data centers. Data centers are secure sites where analysts can access confidential data in a setting that ensures the integrity of confidential micro data. Some agencies have developed routines that allow analysts to submit computer programs remotely across agency firewalls to access confidential microdata. This paper will explore the advantages and issues associated with each type of data access.

Introduction: Access to Statistical Agency Restricted Data

The Federal Statistical Agencies collect a wealth of data on America's society, economy, institutions, and environment. These data are collected to meet specific or general requirements in legislation or the code of federal regulations. The agencies publish estimates from that data in a wide variety of media and formats from specific tabular forms on paper to interactive query systems on the Internet. There is routine reporting of statistics which accumulate over time into time series monitoring trends. There are special studies of topical interest. And there are detailed analyses published in scientific journals. Initial publication could be a press release followed by bulletins that present much statistical data and analysis. Usually there are still many additional possible tabulations and analyses not published due to the lack of resources within the agency.

Outside the federal statistical agencies there are many institutions with interests in science and public policy that have resources to support tabulation and analysis of data produced by these agencies. And in a free society there is compelling interest in making data available to the public for analysis and publication. Indeed, most of the federal statistical agencies devote considerable resources to the preparation and publication of public use microdata data files (PUMS). At this point, however, the agencies encounter conflicting requirements. Data rich PUMS files contain records representing individuals or establishments. The detailed attributes on these records include some of the complex characteristics of individuals or establishments that make them unique, and thus create the possibility that someone might recognize an individual or organization in the data file. But, many of these data are collected under pledges of confidentiality. In some cases, agency employees releasing identifiable information are subject to severe legal penalties.

The agencies use statistical disclosure control techniques to protect individual identification. These techniques involve data modification or partial suppression to avoid the release of data so

detailed that individual respondents can be identified. Agencies have policies and rules governing the publication of statistical tables and analyses. For instance, in publishing total counts or amounts, agencies inspect tables to be sure that at least 3 organizations contribute to the total and that no one organization contributes more than one-half. This restriction makes it impossible for one organization to deduce a competitor's response from a published table. In publishing PUMS agencies remove obvious identifiers and use statistical disclosure control techniques to protect the identity of individual respondents. These techniques involve data modification or partial suppression, such as coding continuous amounts into categories and grouping all extreme cases into cells less than or greater than cut-off amounts.

Threats to Data Confidentiality

Modern computing power plus the information explosion has increased the vulnerability of federal statistical agency data to re-identification. Let's examine this issue in more detail.

There is an unprecedented growth in the size, detail and variety of data collections as computer technology and disk storage space become increasingly affordable. Latanya Sweeney has summarized this as a tendency to collect more, collect specifically, and collect it if you can. Although federal statistical agencies are probably less likely to respond opportunistically in the current environment, they are certainly not immune, and some of our greatest achievements of recent decades are part of this trend. For instance, in 1960 our system of economic statistics was mainly producing national estimates; now we are getting estimates for some statistics down to the county level. This is the result of increases in the size of data collections like the Current Population Survey.

An example of an increase in detail is the birth certificate. The fields on birth certificates in the mid 1900's included little beyond fields identifying the child, the child's parents, and the place of birth. There were a few demographic and medical fields for birth order, weight, and health status. Today, in addition to the basic information collected in the mid 1900's, birth records include additional information on parents such as their education and place of residence at the birth date, on the mother's health, risk factors and health care, and on the infant's health and delivery. Eight States have open vital records files and twnty-five have restricted access procedures.

An example in the private sector is storage of customer transactions in supermarkets utilizing loyalty cards. Food Marketing Institute reported in 1998 that 6 out of 10 supermarket companies collect or plan to collect detailed information on consumer purchases compared to 3 out of 10 in 1993. In terms of supermarket collection a consumer can opt out by not participating in a loyalty card program but can not opt out of mandatory government programs such as birth certificate records.

On the collection side there is no doubt that we are moving toward an environment where society could collect and store data on all persons; one of the fields added to the birth certificate is a check-off box requesting a social security number for the infant. On the access side technology is making the transfer of data very easy. In the past to view a paper record you had to travel to the record repository or have someone copy and mail the information. Computers and public use

files made data available to select individuals with programming skills and access to computer systems. Today the power of the personal computer, software and the internet permits personal data to be transmitted across the street or around the world. CD-rom and DVD technology make inexpensive storage and distribution widely available and reduce access time. Distance has been replaced by the speed of one's connection to the internet; and there is no reason to believe that this will long remain a limiting factor.

Today on the internet it is easy to identify data bases that have detailed personal information about people, companies, etc. The ability of a user to take public statistical agency datasets and link them with other easily available data limits the amount of detail that can be included on PUMS files.

Making Micro Data Available: Restricted Data, Tradition Methods, PUMS

Agencies also release public use datasets for researchers to further analyze on their own. When agencies release public use datasets for researchers to further analyze the amount of detail that can be released must be limited. Obvious identifiers such as name, address, and social security number are not released. Sensitive data elements such as annual salary are typically top coded or only reported in fairly wide bands. Geographic detail is often restricted at areas that have population totals over 100,000, 250,000, or even greater.

Economic establishment data are never released as public use datasets. Geographic identifiers on demographic and social statistics must be suppressed or aggregated to levels that limit the analysis possible. No identifiers are included on PUMS files that would enable a researcher to link agency data with other data.

Bureau of Labor Statistics (BLS) releases special CPI data sets for researchers when requested. These sets often include longitudinal prices within establishments. We have found that the interaction between variables on these datasets must also be evaluated to ensure that a knowledgeable person can not isolate out individually identifiable respondent information.

For example, the National Center for Health Statistics (NCHS) is pleased to offer downloadable public-use data files through the Centers for Disease Control and Prevention's (CDC)/FTP file server. The web site offers the following documentation of downloading PUMS files:

> Users of this service have access to **data sets**, **documentation**, and **questionnaires** from NCHS surveys and data collection systems. Downloading instructions are available in "readme" files.
> Public-use data files are prepared and disseminated to provide access to the full scope of the data. This allows researchers to manipulate the data in a format appropriate for their analyses. NCHS makes every effort to **release data** collected through its surveys and data systems in a timely manner. Descriptions of NCHS data systems and activities are found in the section **Surveys and Data Collection Systems**. Public-use data files that are not listed below can be obtained through other sources. Ordering instructions and the various formats available (e.g., CD-ROM and data tape) are provided in the **Electronic Products** web pages. Users of NCHS public-use data files must comply with **data use restrictions**

to ensure that the information will be used solely for statistical analysis or reporting purposes.

Since the statistical agencies can only produce a limited amount of potential outputs, the full potential of these data are not realized. One way of satisfying both concerns, the desire of researchers to have access to such files and the desire to prevent disclosures, is for the agency or research organization to release files under highly controlled conditions. This article will explore four methods of restricted access procedures that are used to allow researchers to access confidential data:

- Licensing Agreements
- Research Fellowships and Post-Doctoral Programs
- Research Data Centers
- Remote Access.

The later two methods will be explored in detail in this paper.

Licensing Agreements

A licensing agreement is a formal agreement that permits confidential microdata to reside on a researcher's personal computer in their home institution. These agreements are formal legal documents between the agency and the host organization that specify the conditions under which the specific data set licensed may be used and the penalties for violation.

There are several common themes that run through the licensing agreements.

The principal investigator (PI) must demonstrate that the data are required for research; i.e., public use data, if it exists, are not adequate. The goals of the research that require non-public data must be stated in the application. The licensor must approve the goals of the research before the application process can proceed.

License agreements specify which people in the licensee's institution must sign the form. For an academic department it is typically a Dean and not the department chairman.

The agreement also includes a statement concerning which law(s) protects the data (e.g., Privacy Act of 1974). The PI must supply a list of names of people who will be authorized to use the data. Those people must be informed of their responsibility not to share the data with people outside the group. The PI must indicate the group's experience, if any, with handling other licensed datasets.

A data security program must be developed and implemented. The licensee's institution must allow inspections of the area where the data are used and stored.

Inspections of licensee's institution are used to enforce the data security program and access restrictions. The inspections can be unannounced. Penalties for violations of aspects of the agreement are listed on the form (e.g., denial of use of other data from the licensor, fines, prison

terms, etc.). There is a requirement that no attempt will be made to determine the identity of respondents. In general, the licensee is not allowed to link the licensed data to other microdata files.

Articles, reports, and statistical summaries generated from the data must be reviewed by the agency before they are published or otherwise communicated. The results must adhere to the agency's disclosure limitation practices (e.g., all non-zero cells in a publicly released table must represent some minimum number of respondents).

Some examples of datasets released under licensing agreements include: National Center for Educational Statistics (NCES)'s Schools and Staffing Survey and The Early Childhood Longitudinal Study; BLS's Census of Fatal Occupational Injuries and The National Longitudinal Survey of Youth; and National Science Foundation (NSF)'s Survey of Doctorate Recipients and Survey of Earned Doctorates.

To date, statistical agencies have found no flagrant violations of the licensing agreements that would warrant requesting the U. S. Department of Justice (DOJ) to prosecute an individual. The question to ask is: Would DOJ consider a confidentiality breach a serious enough offense to prosecute? If not, what message would we be sending to our respondents about the seriousness of the stewardess of the data entrusted to us?

Fellowships and Post Doctoral Programs in Principal Statistical Agencies

Research Fellowships and post-doctoral programs provide unique opportunities for researchers to address some of the complex methodological problems and analytic issues relevant to agency's programs. Fellows and Post-doctoral candidates conduct research in residence at an agency, use agency data and facilities, and interact with agency staff. They adhere to the same confidentiality agreements as regular employees.

Research fellows have to have a recognized research record and considerable expertise in their area of proposed research. The American Statistical Association (ASA) administers the ASA/NSF Research Fellowship Programs, with some support from the NSF for three Federal statistical agencies: the Bureau of the Census (BOC), the BLS, and the NCES. The ASA also administers a Research Fellowship Program for the NCHS and the Bureau of Economic Analysis (BEA).

Restricted Data Access: Research Data Centers

Research Data Centers (RDCs) are secure facilities designed to provide outside researchers access to confidential microdata files. Initially these facilities have been located only at an agency's headquarters. After gaining sufficient experience with these centers agencies may expand them to additional locations. The BOC, for instance, has expanded its RDC program to various sites around the country. RDCs are both physically and electronically separated from agency's central data stores and routine operations.

After an agency has decided to create a center by gaining agreement from within and outside, decisions have to be made about which data will be made available for access. These decisions include the survey files that will be available for analysis and the data elements collected that will be made available. Some files, such as Internal Revenue Service tax files, may be considered too sensitive to allow non-agency personnel access. Permissions may need to be need to be obtained from survey sponsors (some of which may be in other government departments), providers of administrative data underlying the agency's programs, and possibly higher levels within the agency's Department (such as departmental legal offices). Files should have adequate documentation on definitions, data fields, etc.

The specific details that make RDCs possible varies from agency to agency subject to the legal protections of data. Access to certain sensitive identifiers such as name, address, social security number may not be allowed. Outside researchers might have conditions placed on use that are more restrictive than internal staff. The BOC has authority to make researchers special sworn employees, which subjects them to the same penalties as agency employees for confidentiality breaches. Other agencies do not have this authority and must, as a result, be more restrictive in making data available. Agencies might restrict access for the sake of research only or to projects that generate specific benefits to the agency's programs; this is one of the requirements at the BOC, but not at NCHS.

In choosing site locations care must be exercised to ensure that the selection process is fair. Solicitation announcements should be made in the federal register in addition to distribution to likely candidate organizations. It might be advisable choose the sites with a partner such as the NSF as the US BOC did. The evaluation process should be fair and objective. As RDCs impose considerable costs on the agency, and the agency must decide which options to use to recover the costs associated with RDCs. Costs can be recovered by charging researchers directly or charging the host organizations which can recover their costs by charging laboratory fees. The BOC and the NCHS charge researchers directly at headquarters. BOC charges hosts for remote sites.

The RDCs must be secure facilities not only physically but also procedurally. All materials researchers remove from the facilities must be reviewed for confidentiality. The computer facilities must have no network or internet links to or from the outside and the "A" drive and/or other write media disabled. The site must have an on-site employee or contractor who is trained in security and the datasets.

The NCHS has as RDC only at its headquarters while BOC has remote locations in addition to its Washington, D.C. headquarters. The NCHS RDC is a secure monitored facility where

external researchers may be allowed access to internal restricted data files for approved projects. Restricted data files are those that contain information, such as lower levels of geography (e.g., state, county, or Census tract), but do not contain direct identifiers (e.g., name or social security number). Restricted data files may be used in the RDC by researchers wishing to control for geographic area in their models or they may be used to merge additional data onto the NCHS collected data files for enhanced analyses (e.g. The NCHS contextual data file.) To gain access to the NCHS RDC researchers must follow the strict procedures that govern the use of the RDC:

- researchers must submit a research proposal
- no materials may be brought into the RDC
- no materials, printed or electronic may leave the RDC without a disclosure review
- researchers must sign a Researcher Affidavit of Confidentiality
- the RDC is open only when staff are available for supervision
- use of the RDC is subject to space availability, consistency with the NCHS mission and
- the feasibility of the proposed project.

Except for very unusual circumstances, researchers are not allowed access to files with direct geographic identifiers. Should a researcher request an NCHS data file merged with external data, RDC staff will merge the files then remove the geographic identifiers leaving the researcher access to a files that consists of the NCHS data merged with the additional data. Should the researcher need clustering variables to stratify on geography, RDC staff will construct a set of dummy geographic indicators.

Expanding the number of research data centers beyond agency headquarters has been limited by the expense of developing and maintaining a center and by the difficulty of meeting confidentiality restrictions. Even recognizing that user fees might recover certain costs, everything isn't recouped. There are non-center costs of developing survey documentation, creating center files, training staff on file structure and data limitations, replacing on-site staff, maintaining equipment, etc. And there are issues in management and organization. For instance, NCHS' confidentiality law forbids the public release of confidential data and thus requires that an RDC be staffed by Center employees. Regardless of the staffing, an authority structure has to be created that maintains and enforces agencies' culture of confidentiality.

Restricted Data Access: Remote Access

For many researchers, working at an RDC is a burden because of travel away from his/her host institution. Remote access overcomes, almost, the expense and inconvenience of distance. With remote access researchers outside the statistical agency submit analytical programs through e-mail or the internet to an RDC to run on RDC computers storing confidential microdata files. Here, too, many decisions need to be made. Decisions need to made on the languages that will be supported, medium to be used to submit the programs and review procedures for the output generated. Usually, remote access in not a method that can produce tabulations not previously released.

At NCHS SAS was chosen as the analytic language because it is in wide use and is sufficiently well structured that an automated scanning system could be used. A number of functions

available in SAS have been disabled because they are capable of producing output that present an unreasonable risk of disclosure. These commands might result in a case listing or produce unstructured output that cannot be inspected by the system. The current NCHS remote access system operates by e-mail but an internet-based system is under development and testing. The internet-based system offers a user-friendlier interface and is capable of improved turn-around time.

The RDC staff will construct a dummy data file configured exactly like the real data (univariate distributions are the same, variable locations and lengths are the same, and paths are the same) that the researcher can use for developing and debugging programs prior to sending them to the remote access system. The use of the dummy data file results in fewer iterations on the remote access system thus increasing overall efficiency. The remote access system operates entirely automatically: the system scans the e-mail for arriving computer programs, validates the user, scans programs for forbidden commands, verifies that programs are not trying to access unauthorized data files and, if no problems are found, executes the program against the real data. After execution, the system scans the analytic output generated by users' program for disclosure problems. Questionable output is routed to an RDC staff person for manual resolution. Users can submit requests to the remote access system 24 hours a day although output is only returned during normal working hours because staff randomly spot check the system to ensure that the system is working properly in all respects. Generally users receive their output within a few hours after submitting their e-mail.

Issues in Making Data Available

There are various laws governing confidentiality of data in the federal statistical system. BOC, NCHS, NCES, and Bureau of Transportation Statistics (BTS) each have agency-specific laws specifying the protection of their data. These laws, as illustrated above, are not consistent with each other. Other statistical agencies are covered by more general provisions in exemption B4 of the Freedom of Information Act (FOIA), and the Trade Secrets and Privacy Acts.

Following the various laws, the various agencies have various policies. There is a lack of uniformity in policy across the agencies. Instruments such as licensing which are available to one agency are not available to another. Each agency has to develop procedures customized to their own data and their own legal environment to protect their data and to respond to requests for access. This inhibits the development of protection policies by making it more difficult for agency officials to find common ground either for discussion and policy development or for actual cooperation in the creation of institutions like RDCs. The differences in the legal context of institutions is one reason why it is that the BOC and NCHS have developed RDCs while NCES has developed licensing agreements and the BLS has limited its access program to IPA and fellowship awardees. These differences also mean that the administrative and legal means of enforcement differ across agencies. Because of this variety any one agency has less relevant legal experience and the general legal environment for protecting statistical data is more uncertain than it might be.

The variety of laws governing various statistical agencies also inhibits cooperation among statistical agencies at levels other than policy making. Some examples of this are well-known.

Agencies are prevented from sharing some data on sampling frames, for instance, with the result that one agency is unable to take advantage of advances within another agency, inconsistent data sets are created, and survey costs are increased. Agencies are also limited in their capacity to share data for research purposes. In this case the scientific community and the public are denied the benefit that might flow from linking data across agencies.

The legal restrictions on sharing data also limit the ability of the statistical agencies to share RDC resources. BOC employees or special sworn BOC agents can only view Census Bureau data. Thus if BOC data were located in another agency each RDC staff member would have to be a sworn Census agent and ensure all researchers met the BOC restrictions before gaining access to the data. With each agency having its own legal requirements, an RDC that has to maintain different procedures for different agencies becomes unwieldy.

The public is rightly concerned about the capacity to link data, but the complex legal situation does not facilitate the statistical agencies efforts to explain the risks and protections to the public. Public opinion research shows that the public is skeptical of the government's promises to protect privacy and cynically believes that there is wide-spread data sharing among agencies. The statistical system, which institutionally is committed to protecting respondents, is not getting credit for its position while the public is not getting the benefits of data sharing, of which it thinks it is bearing the costs. This is a lose-lose situation.

The public is not alone in its concern that the confidentiality of statistical data is increasingly threatened. This conference is evidence of concern within the statistical community. As mentioned above the threats to confidentiality are increasing. These threats, however, are but dimly perceived. There has not been very much research focused on the resources available to someone attempting to reidentify entities on PUMS. Even the elementary strategies a data intruder might employ have only been superficially explored. These studies have shown that certain data sets do have limited vulnerability, and that there are data resources that an intruder might use. That is, demonstration projects have shown that in certain files persons targeted because they had met rare criteria might be identified through matching these rare criteria in other publicly available data sets. These studies suggest that there is a need to review and catalog the growing accumulations of data and evaluate them from the perspective of their potential value to a data intruder.

Efforts of the Federal statistical system on detecting a fixed disclosure risk are ad hoc. Problems are fixed as issues are raised. However, there are not many efforts by the agencies in the statistical system to systematically test their PUMS against as many data sets as publicly available as research for identification risk.

For example, recently the BLS was concerned that the National Longitudinal Public Use Datasets were vulnerable for reindentification using birth records. BLS contracted with a researcher to see if he could identify individual respondent data from birth records. The researcher used Massachusetts records along with birth information on the file to verify that with considerable expense it is possible to reidentify some records. BLS decided to suppress detailed birthdate information to ensure adequate protection to the data. However, we need a program to

study all the variables with all publically available datasets to ensure no undetected problems exist.

Research into the vulnerability of published data to reidentification will also support a growing stream of research into techniques of disclosure limitation. The purpose of this stream is to produce techniques that statistical agencies can use to raise barriers to reidentification. This research is important because lacking proven disclosure limitiation techniques statistical agencies will be placed in the unhappy situation of having to withhold data sets from surveys that once were published. That is, rather than continue to expand the public availability of data, agencies will have to retrench and put more of their data under access restrictions such as RDCs or remote access.

The most commonly applied techniques of disclosure limitation in microdata files, recoding schemes and data swapping, are applications of pragmatic, ad-hoc methods. Statistical research has, at this point, largely described the statistical properties of these methods. This research has also defined the problem in statistical terms and established methods for evaluating disclosure limitation techniques. With this as a foundation there is a new stream of research emerging into new methods based on statistical theory. A great deal of work needs to be done in this area, however, before this research produces results with practical application.

There is a continuous demand for more information and more detailed information. In responding to this demand agencies are exploring new ways of producing tables and publishing data using CD-ROM and internet technologies. They are also discovering some of the limitations of existing methods of disclosure limitation in published tables. This is another area in which pragmatic and ad-hoc methods have been analyzed with statistical theory and theoretically motivated methods are beginning to emerge. There is slim hope that these methods can satisfy users demands for information, but there is the greater possibility that these methods can be applied in automated systems such as remote access to restricted data and internet query systems like the Bureau of the Census American Factfinder system.

One last area where research is needed is statistical disclosure in models. Although statistical models generally are not sufficiently precise to lead to the statistical disclosure of confidential information, tables can, in fact, sometimes be expressed in statistical models that then inherit the same problems of potential disclosure inherent in the tabular form. Little research has been done on the vulnerability faced by statistical agencies on allowing researchers to publish intricate models. However, most research on restricted data involves publishing models.

For example, suppose one fits a simple regression model of a dependent variable against three independent variables where the model fit of the independent variables with a dependent variable is exceedingly high. Suppose in a population there exist only one entity with a specific set of values on those independent variables. Then it is possible via the model to determine the exact value of that dependent variable fairly closely.

Another issue with models that needs exploration is the risk to disclosure of sensitive dependent variables using readily available micro data that can be applied to the model's independent variables.

Conclusion

We have explored in this paper four methods Federal statistical agencies use to allow researchers access to confidential micro data: PUMS and restricted access methods. These methods have been devised to allow researchers access to the richness of statistical agency data for further analysis than the agencies can do themselves. It also opens up possibilities for re-analysis using a different approach. That builds up credibility for analysis performed by the statistical agencies.

PUMS have been produced by the agencies for demographic statistics for years. However, the richness of data found on the Internet has shown us the vulnerability of re-identification is a real threat. Ad hoc adjustments have been made. However, we need to consider a systematic review of all PUMS by all agencies producing them for disclosure risk. PUMS for economic data is not a viable due to our inability to minimize disclosure risk while providing a useful file for analysis.

The power of the PC and Internet has allowed statistical agencies the ability to set up restricted access procedures: either remote data centers or remote access. However, these efforts are done by each statistical agency independently. We need to consider setting up a one-stop shopping RDC for access to sensitive research files like FEDSTATS for published series. This will require confidentiality legislation that will give the statistical agencies uniform laws to grant special sworn status to their data. Here too much work is needed. Models proposed by researchers to be published are usually assumed safe and not given a lot of disclosure review. Are they really safe?

References

Doyle, Lane, Theeuwes, Zayatz (2001), Confidentiality, Disclosure and Data Access—Theory and Practical Applications for Statistical Agencies, North-Holland

Massell, Paul, Overview of Data Licensing Agreements at U.S. Government Agencies and Research Organizations, CDAC Paper

Jabine, Thomas B.(1993), "Procedures for Restricted Data Access," *J. Official Statistics*, vol. 9, no. 2, pp. 537-589.

Massell, Paul B.(1999), "Review of Data Licensing Agreements at U.S. Government Agencies and Research Organizations," paper presented at the Workshop on Confidentiality of and Access to Research Data Files, sponsored by the Committee on National Statistics (CNSTAT), Washington, D.C.

Massell, Paul B., Laura Zayatz, (2000), "Data Licensing Agreements at U.S. Government Agencies and Research Organizations," *Proceedings of ICES-II (International Conference on Establishment Surveys).*

George.T. Duncan, Thomas B. Jabine, Virginia A. de Wolf (eds.), *Private Lives and Public Policies,* National Academy Press (1993), in "Chapter 6 : Technical and Administrative Procedures," pp. 141-179.

Jabine, Thomas B., "Procedures for Restricted Use Access." *Journal of Official Statistics*, 9:2, 1993, pp. 537-589.

National Center for Education Statistics, "Restricted Use Data Procedures Manual."

Reznek, Arnold., Joyce. Cooper, and J. Bradford Jensen. "Increasing Access to Longitudinal Survey Microdata: the Census Bureau's Research Data Center Program." *American Statistical Association 1997 Proceedings of the Section on Government Statistics and Section on Social Statistics.* Alexandria, VA, 1997, pp. 243-248.

Sweeney, Latanya, Information Explosion in Doyle, Lane, Theeuwes, Zayatz (2001), Confidentiality, Disclosure and Data Access—Theory and Practical Applications for Statistical Agencies, North-Holland. Pp 43-74

Alonso, William and Paul Starr (eds). 1987. "The Politics of Numbers." New York: Russell Sage Foundation.

Singer, Eleanor, Public perceptions of confidentiality and attitudes toward data sharing by federal agencies in Doyle, Lane, Theeuwes, Zayatz (2001), Confidentiality, Disclosure and Data Access—Theory and Practical Applications for Statistical Agencies, North-Holland. Pp 341-70

Greenia, Nick, J Bradform Jensen and Julia Lane, Business perceptions of confidentiality, in Doyle, Lane, Theeuwes, Zayatz (2001), Confidentiality, Disclosure and Data Access—Theory and Practical Applications for Statistical Agencies, North-Holland. Pp 395-429