Disclosure Limitation Techniques for Tabular Data[†]

Joe Fred Gonzalez, Jr.
National Center for Health Statistics, Office of Research and Methodology
3311 Toledo Rd., Hyattsville, MD 20782

Abstract

In order for the U.S. Federal Statistical System to collect and release reliable and valid data, it must gain the cooperation and trust of its survey respondents (persons or establishments). To achieve this, federal statistical agencies must pledge confidentiality (under penalty of law) to protect respondents' data prior to collection and public release. Survey data released to the public are usually in the form of microdata and tabular data. This paper will focus on techniques for disclosure avoidance of two-dimensional tabular data. Tabular disclosure limitation techniques that will be presented are: complementary cell suppression, minimum-distance controlled rounding, unbiased controlled rounding, controlled rounding subject to subtotals constraints, and controlled tabular adjustment (CTA). Before and after perturbation results will be compared.

Key Words: confidentiality, SDL, avoidance techniques

1. Introduction

The National Center for Health Statistics (NCHS) collects, compiles, and publishes general purpose vital and health statistics which serve the needs of all segments of health and health related professions. The success of the Center's operations depends upon the voluntary cooperation of States, of establishments, and of individuals who provide the information required by the Center programs under an assurance that such information will be kept confidential and can only be used for statistical purposes (1). The NCHS operates under the authority and restrictions of *Section 308 (d) of the Public Health Service Act* (2) which provides in summary that no information obtained in the course of its activities may be used for any purpose other than the purpose for which it was supplied. This collected information may not be published or released in a manner in which the establishment or person supplying the information or described in it is identifiable unless such establishment or person has consented. In addition to *Section 308 (d) of the Public Health Service Act*, the E-Government Act of 2002 (Public Law 107-347) was signed into law on December 17, 2002. Title V of the E-Government Act is the Confidential Information Protection and Statistical Efficiency Act (CIPSEA). CIPSEA's primary purposes are to protect information collected for statistical purposes from improper disclosure and to ensure that the information is not used for nonstatistical purposes.

When does statistical disclosure occur? Statistical disclosure occurs when the release of a statistical data product enables a third party to learn more about a respondent than the

[†] The findings and conclusions in this paper are those of the author and do not necessarily represent the views of the National Center for Health Statistics, Centers for Disease Control and Prevention.

third party had already known (Dalenius). A major responsibility of the NCHS is the protection of identifiable data collected from survey respondents (persons or establishments) directly or indirectly. Prior to release of public use files, data that could be used to identify a respondent are perturbed, suppressed, or removed from microdata files. The other mechanism for statistical disclosure is the possible identification of individuals or establishments via cell counts in tables.

Possible tabular data disclosure may occur in various ways. For example, respondents may be identified directly from small cell sizes (1, 2, 3, 4) in categorical data, or via magnitude data (number of events, such as hospital admissions or discharges where each respondent can contribute unequally to each cell in a table), respondent contributions to heavily concentrated cells may be closely approximated by the cell value. One possible solution is to round cell counts to base 5, subject to the following constraints: multiples of 5 remain fixed; non-multiples of 5 round to one of the two adjacent multiples of 5; rounded table is additive; rounding procedure is random (unbiased). Indirect tabular statistical disclosure is also possible in tables through manipulation of additive tabular relationships between cell values and totals, e.g., manipulating rows and column totals in a two-dimensional table.

This paper will focus on techniques for disclosure avoidance of two-dimensional tabular data. The tabular disclosure limitation techniques that will be presented are: complementary cell suppression, minimum-distance controlled rounding, unbiased controlled rounding, controlled rounding subject to subtotals constraints, and controlled tabular adjustment (CTA). The statistical disclosure limitation (SDL) software discussed here which can invoke each of the five tabular disclosure limitation techniques was developed, under contract with the NCHS, by OptTek Systems Inc. Hereafter, each of the above tabular disclosure limitation techniques will be referred to as software functions. Each section which follows will describe and apply each of the five software functions to Table 1 below which contains the original cell counts which were randomly generated and do not represent real data values.

2. Cell Suppression

A multiple-cell suppression technique developed by Cox (1995) is used as the cell suppression function in the software. This software function hides from publication the values of all cells representing direct disclosure of confidential data on individual respondents (the *disclosure cells (blue)).* As a consequence, a sufficient number of selected nondisclosure cells (the *complementary cells(red))* are also suppressed to ensure that a third party cannot reconstruct or narrowly estimate confidential respondent data by manipulating linear relationships between released and suppressed table values.

The challenge of the cell suppression function is to select complementary suppressions that provide sufficient disclosure protection for the cell counts while minimizing the amount of information lost due to suppression. Information loss for tables is typically measured as:

- number of cells suppressed,
- total value suppressed, and
- total percent of value suppressed or other functions such as the total of of logarithm of one plus value suppressed.

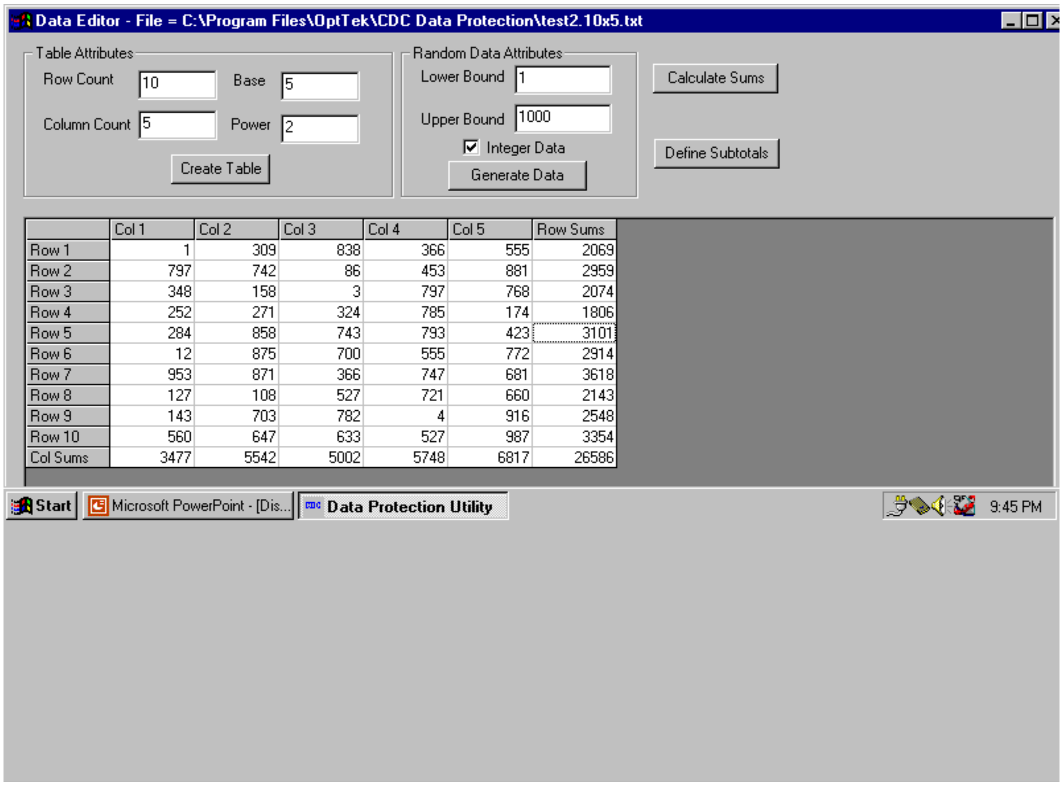
An objective function (or cost function) is formed to minimize one of the above three information losses. Then, the cell suppression is carried out by minimizing the objective function subject to certain constraints, such as row or column marginal totals. The choice of objective function is an important one for the data provider:

- the SDL software can produce optimal solutions for different objective functions, but the resulting solutions can be quite different.

- It is important that the provider, through prior experimentation or established policy, understand the qualitative difference likely to result from the specification of a particular objective function.

The cell suppression function used in the SDL software is based on mathematical networks which offer theoretical and practical advantages (reduced polynomial computing time, instead of exponential time) when optimizing certain objective functions. A mathematical network is a specialized linear program defined over a mathematical graph. Table 1 below represents the original cell counts. Each software function will operate on Table 1.

Table 1. Original Cell Counts used as Input for all Software Functions.



Data Editor - File = C:\Program Files\OptTek\CDC Data Protection\test2.10x5.txt

Table Attributes

Row Count 10

Base 5

Column Count 5

Power 2

Random Data Attributes

Lower Bound 1

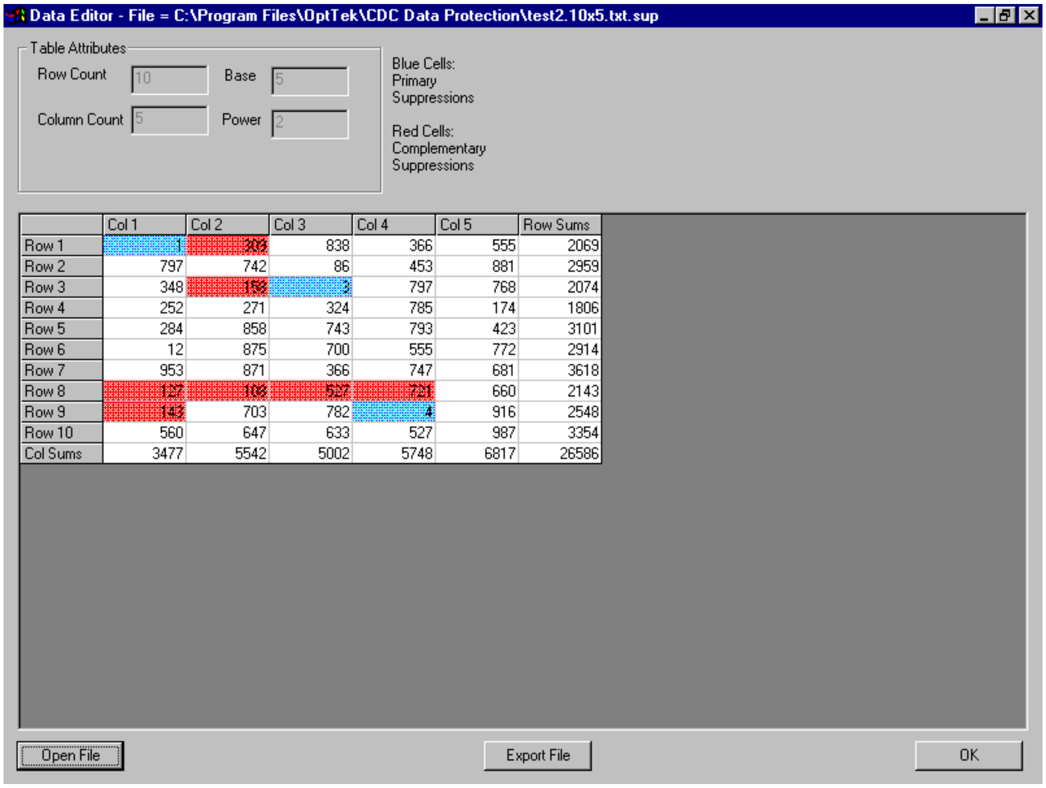
Upper Bound 1000

☑ Integer Data

	Col 1	Col 2	Col 3	Col 4	Col 5	Row Sums
Row 1	1	309	838	366	555	2069
Row 2	797	742	86	453	881	2959
Row 3	348	158	3	797	768	2074
Row 4	252	271	324	785	174	1806
Row 5	284	858	743	793	423	3101
Row 6	12	875	700	555	772	2914
Row 7	953	871	366	747	681	3618
Row 8	127	108	527	721	660	2143
Row 9	143	703	782	4	916	2548
Row 10	560	647	633	527	987	3354
Col Sums	3477	5542	5002	5748	6817	26586

Table 2 below displays the results of applying the complementary cell suppression function to Table 1 (original cell counts). Here, the objective function minimizes the total value suppressed.

Table 2. Results of Applying Complementary Cell Suppression to Table 1 (original cell counts). [NOTE: Primary suppression cells are in blue; Complimentary suppression cells are in red.]



Data Editor - File = C:\Program Files\OptTek\CDC Data Protection\test2.10x5.txt.sup

Table Attributes

Row Count: 10

Base: 5

Column Count: 5

Power: 2

Blue Cells: Primary Suppressions

Red Cells: Complementary Suppressions

	Col 1	Col 2	Col 3	Col 4	Col 5	Row Sums
Row 1	1	309	838	366	555	2069
Row 2	797	742	86	453	881	2959
Row 3	348	158	3	797	768	2074
Row 4	252	271	324	785	174	1806
Row 5	284	858	743	793	423	3101
Row 6	12	875	700	555	772	2914
Row 7	953	871	366	747	681	3618
Row 8	127	108	527	721	660	2143
Row 9	143	703	782	4	916	2548
Row 10	560	647	633	527	987	3354
Col Sums	3477	5542	5002	5748	6817	26586

Unfortunately, there are disadvantages of cell suppression since it limits the utility of tabular data. How do you perform a statistical analysis of a table where some cell counts have
been suppressed?

3. Minimum-Distance Controlled Rounding

Minimum-Distance (or optimal) Controlled rounding is the second function used in the software and is based on the methodology described by Cox and Ernst (1982) and by Causey, Cox, and Ernst (1985). It rounds all entries in a one or two-way tabular array A to integer multiples of a positive integer base B subject to the following requirements:

(1) each entry (a_{ij}) in table A is rounded to an *adjacent integer multiple of B, R(a_{ij})*; that is, an entry a_{ij} is rounded to either *B*[*a/B*] *or B*([*a/B*] + *1*), where [ ] is the greatest integer function, and

(2) the sum of the original values for any row (or column) of table *A* equals the

rounded value of the corresponding row (or column) for the rounded table, *R(A)*. Similarly, rounded values of the row and column totals both sum to the rounded grand total.

Minimum distance (or optimal) controlled rounding can be achieved by presenting this problem as a capacitated transportation problem whose objective function is minimized with respect to the l_p norm, $1 \leq p < \infty$, where the objective function is the pth root of the sum of the pth powers of the absolute values of the differences between rounded and unrounded entries of A. That is,

the objective function to minimize with respect to l_p norm is defined as

$$l_p[R(A), A] = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} \left| R(a_{ij}) - a_{ij} \right|^p \right)^{1/p}$$

where all notation has been previously defined.

- Cox and Ernst (3) showed that this objective function can be expressed as a linear function of appropriate variables, thus defining a linear program.

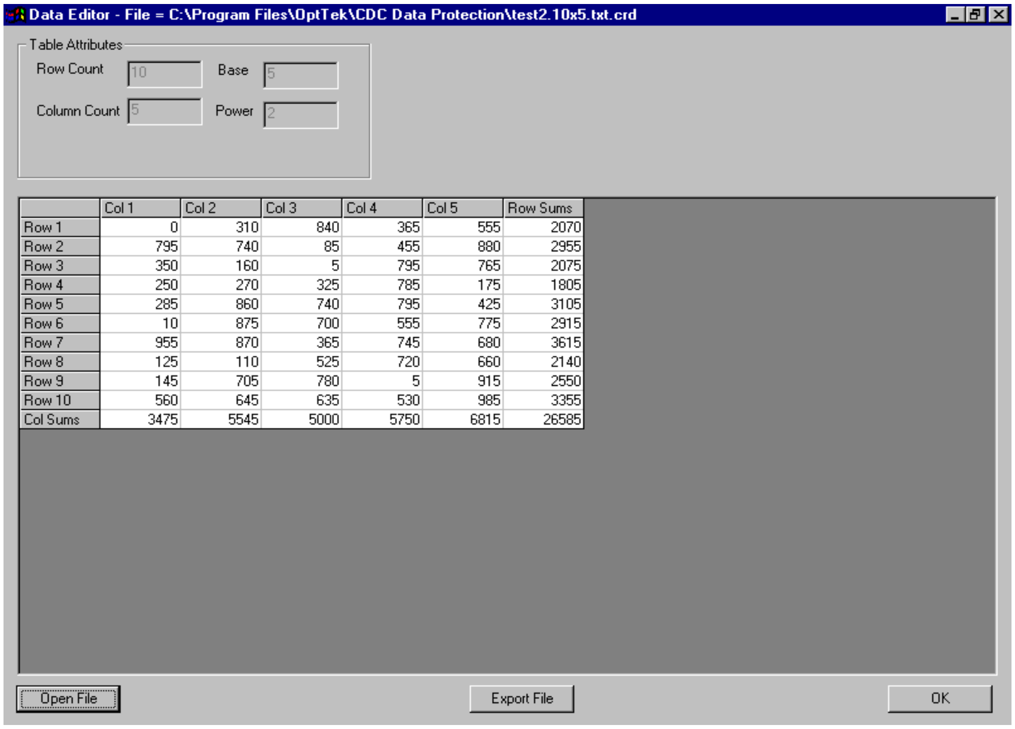
- In general, this function has a geometrical and not a statistical, interpretation (namely, minimum of a corresponding Euclidean distance).

- Unfortunately, minimum-distance controlled rounding cannot be extended to three- or higher-dimensional tables in all cases (Ernst (9)).

The computational time for a minimum-distance controlled rounding solution is dependent on the number of cells in the table that are not multiples of the base, and the number of rows and columns in the table.

- A table with 50 rows and 50 columns was rounded in less than a minute.
- A table with 100 rows and 100 columns was rounded in 24 minutes.
- A table with 1000 rows and 5 columns was rounded in 1 hour and 40 minutes.

Table 3. Results of Applying Controlled Rounding to Table 1 (original cell counts).



Data Editor - File = C:\Program Files\OptTek\CDC Data Protection\test2.10x5.txt.crd

Table Attributes

Row Count: 10 Base: 5

Column Count: 5 Power: 2

	Col 1	Col 2	Col 3	Col 4	Col 5	Row Sums
Row 1	0	310	840	365	555	2070
Row 2	795	740	85	455	880	2955
Row 3	350	160	5	795	765	2075
Row 4	250	270	325	785	175	1805
Row 5	285	860	740	795	425	3105
Row 6	10	875	700	555	775	2915
Row 7	955	870	365	745	680	3615
Row 8	125	110	525	720	660	2140
Row 9	145	705	780	5	915	2550
Row 10	560	645	635	530	985	3355
Col Sums	3475	5545	5000	5750	6815	26585

4. Unbiased Controlled Rounding

Unbiased controlled rounding is the next function that is used in the software and is based on the methodology described by Cox (1987).

The conditions for unbiased controlled rounding are that every entry *a* of A satisfies the following

1. R(a) = B[a/B] or B([a/B] + 1)
2. R(A) is additive.
3. $|$ R(a) – a $|$ < B
4. E [R (a)] = a

In lieu of optimizing with respect to a Euclidean measure of distance, the desired solution preserves original values with respect to the statistical property expectation.

Unbiased controlled rounding test results:

- A table with 50 rows and 50 columns was rounded in a second.
- A table with 100 rows and 100 columns was rounded in 4 seconds.
- A table with 400 rows and 25 columns was rounded in 5 seconds.
- A table with 2000 rows and 25 columns was rounded in 5 minutes and 45 seconds.

Table 4. Results of Applying Unbiased Controlled Rounding to Table 1 (original cell counts).

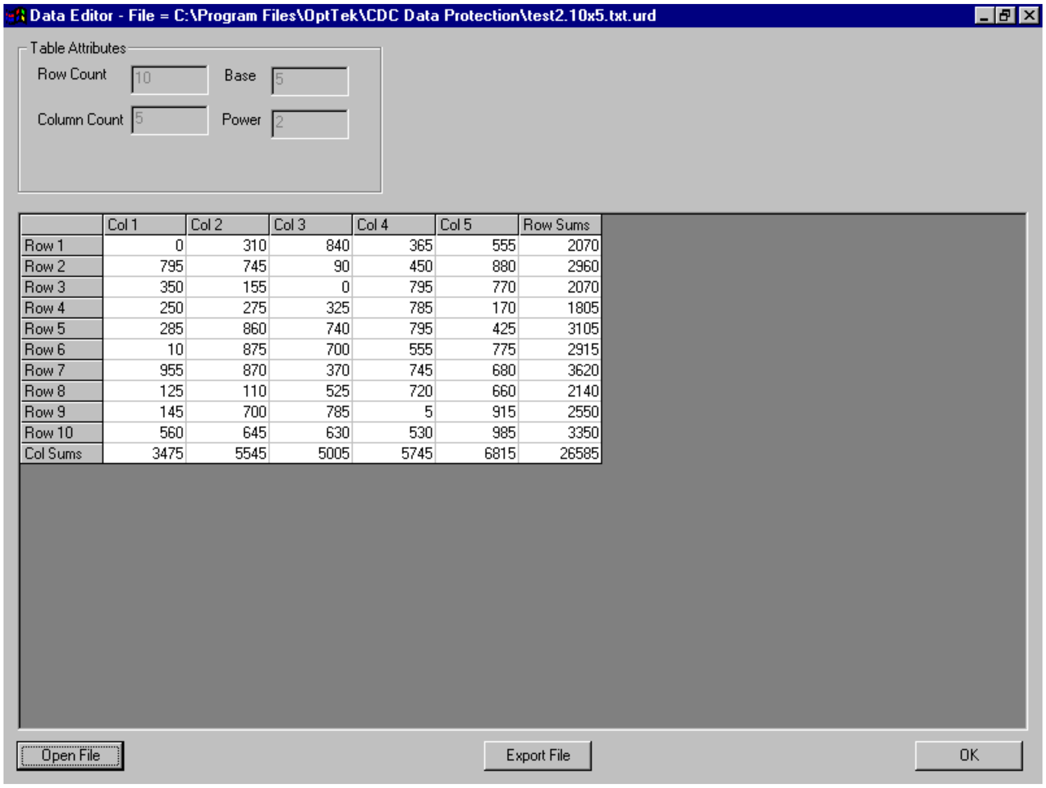

Data Editor - File = C:\Program Files\OptTek\CDC Data Protection\test2.10x5.txt.urd
Open File
Export File
OK

Table Attributes

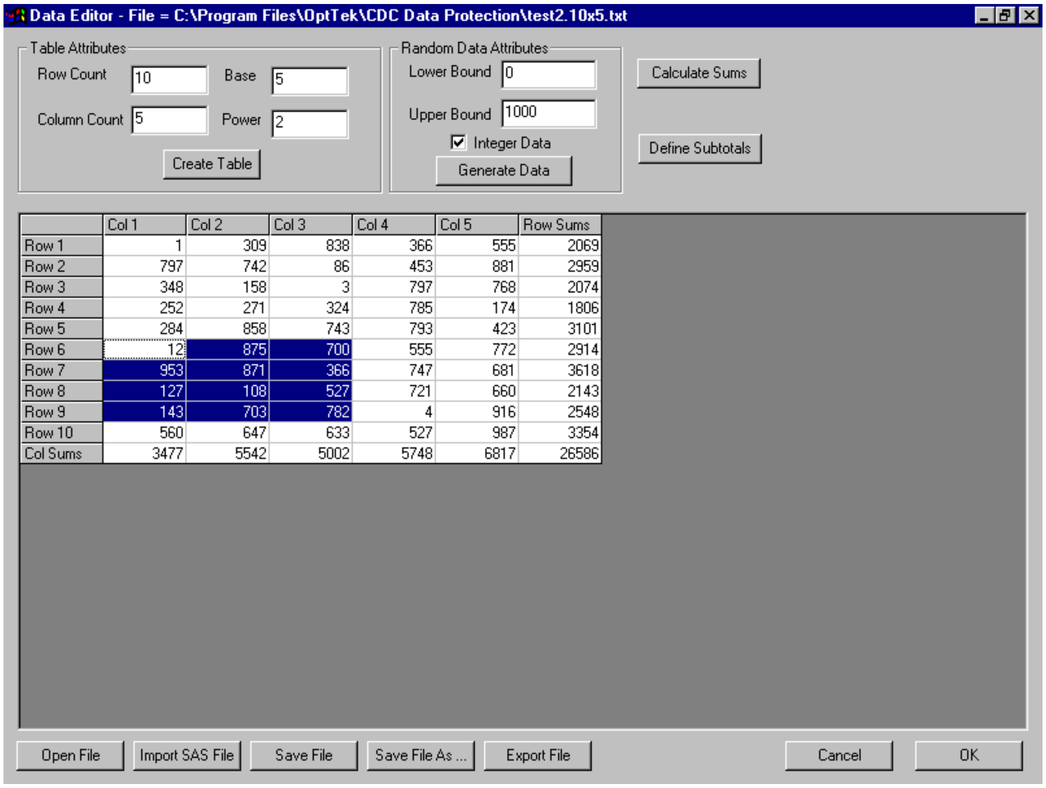
Row Count: 10
Base: 5
Column Count: 5
Power: 2

	Col 1	Col 2	Col 3	Col 4	Col 5	Row Sums
Row 1	0	310	840	365	555	2070
Row 2	795	745	90	450	880	2960
Row 3	350	155	0	795	770	2070
Row 4	250	275	325	785	170	1805
Row 5	285	860	740	795	425	3105
Row 6	10	875	700	555	775	2915
Row 7	955	870	370	745	680	3620
Row 8	125	110	525	720	660	2140
Row 9	145	700	785	5	915	2550
Row 10	560	645	630	530	985	3350
Col Sums	3475	5545	5005	5745	6815	26585

5. Controlled Rounding Subject to Subtotal Constraints

The function, controlled rounding subject to subtotal constraints, is similar to controlled rounding and is based on the methodology described by Cox and George (1987). This function extends that methodology to tables with subtotals along one, but not both, dimensions.

Table 5. Results of Applying Controlled Rounding Subject to Subtotal Constraints to Table 1 (original cell counts).



Data Editor - File = C:\Program Files\OptTek\CDC Data Protection\test2.10x5.txt

Table Attributes

Row Count: 10
Base: 5
Column Count: 5
Power: 2

Random Data Attributes

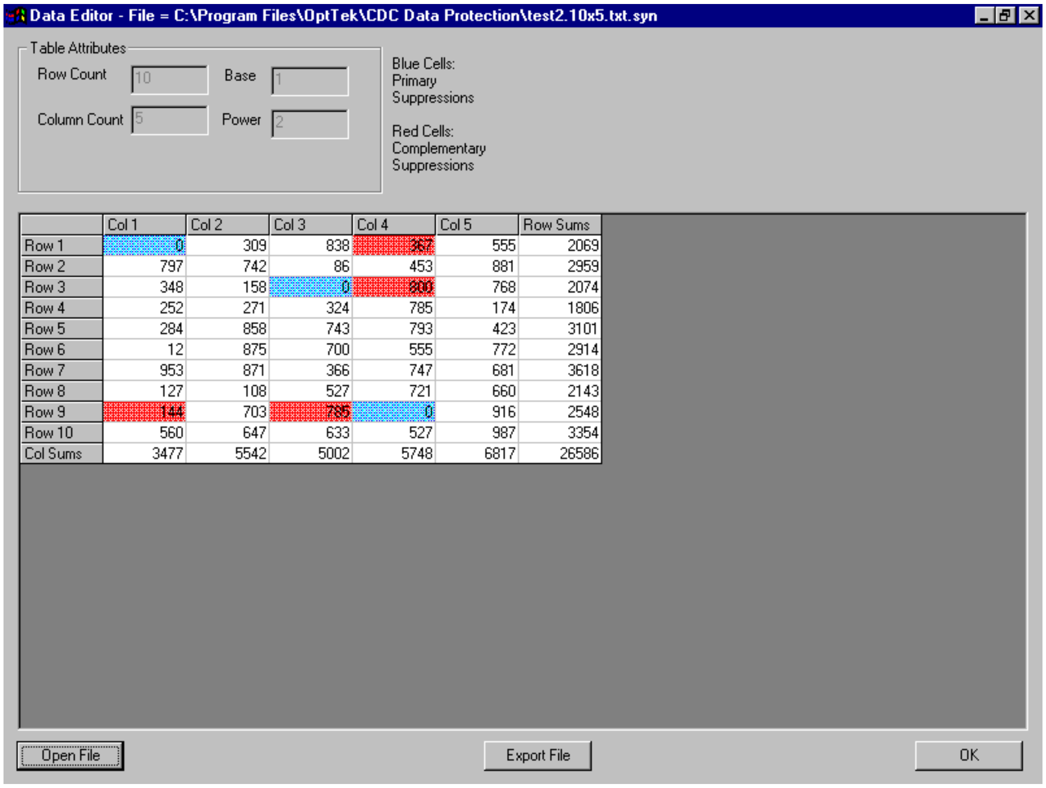
Lower Bound: 0
Upper Bound: 1000
☑ Integer Data

	Col 1	Col 2	Col 3	Col 4	Col 5	Row Sums
Row 1	1	309	838	366	555	2069
Row 2	797	742	86	453	881	2959
Row 3	348	158	3	797	768	2074
Row 4	252	271	324	785	174	1806
Row 5	284	858	743	793	423	3101
Row 6	12	875	700	555	772	2914
Row 7	953	871	366	747	681	3618
Row 8	127	108	527	721	660	2143
Row 9	143	703	782	4	916	2548
Row 10	560	647	633	527	987	3354
Col Sums	3477	5542	5002	5748	6817	26586

6. Synthetic Substitution (Controlled Tabular Adjustment)

Synthetic Substitution (Controlled Tabular Adjustment) was developed by Dandekar and Cox (2002) as an alternative to complementary cell suppression. This procedure uses a threshold rule(s) to determine how cells can be modified. The CTA function replaces the value of each disclosure cell by either of its closest safe values, viz., cell value plus or minus the protection limit, and uses linear programming to make small adjustments to other cells to restore the additive tabular structure. For count data, the safe value is either zero or the threshold n. In Table 6 below, the blue color cells represent the primary (or disclosure) cells, whereas the red color cells represent the complementary cells.

Table 6. Results of Applying Controlled Tabular Adjustment to Table 1 (original cell counts).



Data Editor - File = C:\Program Files\OptTek\CDC Data Protection\test2.10x5.txt.syn

Table Attributes

Row Count: 10
Base: 1
Column Count: 5
Power: 2

Blue Cells:
Primary
Suppressions

Red Cells:
Complementary
Suppressions

	Col 1	Col 2	Col 3	Col 4	Col 5	Row Sums
Row 1	0	309	838	367	555	2069
Row 2	797	742	86	453	881	2959
Row 3	348	158	0	800	768	2074
Row 4	252	271	324	785	174	1806
Row 5	284	858	743	793	423	3101
Row 6	12	875	700	555	772	2914
Row 7	953	871	366	747	681	3618
Row 8	127	108	527	721	660	2143
Row 9	144	703	785	0	916	2548
Row 10	560	647	633	527	987	3354
Col Sums	3477	5542	5002	5748	6817	26586

7. Future Research and Development

The software developed for this project is a tool which features some of the different Statistical Disclosure Limitation (SDL) methods for protecting potential disclosure cell values in two-way tables. One of the goals of this project is to develop production level software that can be embedded into NCHS data analysis activities, for example, analysis conducted in the NCHS Research Data Center (RDC). We expect to incorporate new and improved methods into the software as they become available. SDL research continues at NCHS and elsewhere. As a result of all the SDL research that has transpired, a large body of research (risk vs. utility) has risen in the area of data quality. An example of research addressing data quality and data confidentiality concerns (via CTA) is discussed by Cox and Kelly (11).

References

1. National Center for Health Statistics Nondisclosure Affidavit (unpublished internal document).
2. National Center for Health Statistics website
http://www.cdc.gov/nchs/about/policy/confiden.htm
3. Gonzalez, J.F. and Cox, L.H. (2005). Software for tabular data protection. Statistics in Medicine, 2005; 24:659-669.
4. Cox, L.H. (1995). Network models for complementary cell suppression. Journal of the American Statistical Association 90, 1453-1462.
Cox, L.H. (1996). Addendum. Journal of the American Statistical Association 91, 1757.
5. Cox, L.H. and L.R. Ernst (1982). Controlled rounding. INFOR 20, 423-432.
6. Causey, B.D, L.H. Cox, and L.R. Ernst (1985). Applications of transportation theory to statistical problems. Journal of the American Statistical Association 80, 903-909.
7. Cox, L.H. (1987). A constructive procedure for unbiased controlled rounding. Journal of the American Statistical Association 82, 520-524.
8. Cox, L.H. and J.A. George (1989). Controlled rounding for tables with subtotals. Annals of Operations Research 20, 141-157.
9. SAS Institute Inc., SAS/STAT User's Guide, Version 8, Cary, NC: SAS Institute Inc (1999).
10. Shah, B., Barnwell, B., Bieler, G., SUDAAN Users's Manual, Release 7.0, Research Triangle Park, NC: Research Triangle Institute (1996).
11. Cox, LH, On Properties of multi-dimensional statistical tables. Journal of Statistical Planning and Reference 2003.
12. Ernst, LR. Further applications of linear programming to sampling problems. Technical Report –Census/SRD/RR-89-05. 1989;Washington , DC, US Census Bureau
13. Cox LH., Kelly JP. Ensuring data quality and confidentiality for tabular data. Proceedings of the UNECE/Eurostat Work Session on Statistical Data Confidentiality (Invited Paper), April 2003, Luxembourg; Available:
http://www.unece.org/stats/documents/2003.04.confidentiality.htm