STATISTICS IN MEDICINE
Statist. Med. 2005; **24**:659–669
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2043

Software for tabular data protection[‡]

Joe Fred Gonzalez Jr[*,†] and Lawrence H. Cox

Centers for Disease Control and Prevention, National Center for Health Statistics, Office of Research and Methodology, Hyattsville, MD 20782, U.S.A.

SUMMARY

In order for national statistical offices to maintain the trust of the public to collect data and publish statistics of importance to society and decision-making, it is imperative that respondents (persons or establishments) be guaranteed privacy and confidentiality in return for providing requested confidential data. Consequently, for most survey and census data, disclosure limitation techniques must be applied before the data are ready for public release. For microdata, examples of methods that can be used to identify respondents include directly extracting identifying information from microdata files or indirectly identifying respondents by matching a given file with an external file. For tabular data, respondents may be identified directly from small cell counts or respondent contributions to heavily concentrated cells of magnitude data may be closely approximated by the cell value. Indirect disclosure is 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. Two-dimensional statistical tables are a staple of official statistics. This paper describes a desktop software system that for the first time implements within a single framework four standard disclosure limitation techniques for protecting tabular data in two-dimensional tables: complementary cell suppression, minimum-distance controlled rounding, unbiased controlled rounding, and controlled rounding subject to subtotals constraints, and a fifth, new method: controlled tabular adjustment, and summarizes the five methods. Published in 2005 by John Wiley & Sons, Ltd.

KEY WORDS: statistical disclosure; disclosure limitation; mathematical network

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 programmes under an assurance that such information will be kept confidential and be used only for statistical purposes [1].

[*] Correspondence to: Joe Fred Gonzalez Jr, National Center for Health Statistics, Office of Research and Methodology, 3311 Toledo Road, Room 3121, Hyattsville, MD 20782, U.S.A.
[†] E-mail: jgonzalez@cdc.gov
[‡] This article is a U.S. Government work and is in the public domain in the U.S.A.

Published in 2005 by John Wiley & Sons, Ltd. *Accepted July 2004*

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, and that such 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 legally mandated requirements, the NCHS and other national statistical offices have an ethical responsibility to preserve respondent confidentiality, stemming, for example, from the Code of Professional Ethics of both the American Statistical Association and the International Statistical Institute. It is also a central practical issue for national statistical offices to do so, namely, to maintain high respondent confidence and rates of response and data accuracy. Consequently, a major responsibility of the NCHS is the protection of identifiable data collected from survey respondents, persons or establishments.

Prior to release of public use files, data that could be used to identify a respondent are perturbed or removed from microdata files, that is, files that consist of individual records, each containing values of variables for a single person, business establishment or other unit. Another mechanism for statistical disclosure is the possible identification of individuals or establishments via tabular data. For tabular data, respondents may be identified directly from small cell counts in categorical data or, in magnitude data (number of events, such as hospital admissions or discharges where each respondent can contribute unequally to each cell), respondent contributions to heavily concentrated cells may be closely approximated by the cell value. For example, it may be possible to identify a small hospital in a particular region in the United States based on the small number of reported admissions or discharges. Similarly, the same type of possible disclosure could result with a large hospital having a large number of reported admissions or discharges. Indirect disclosure is 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). Two-dimensional statistical tables are a staple of official statistics. The NCHS has sponsored the development of disclosure limitation software for two-dimensional tables by OptTek Systems, Inc. This paper will describe features of the software and the underlying methods including its different functions: cell suppression; minimum-distance controlled rounding; unbiased controlled rounding; controlled rounding subject to subtotal constraints; and controlled tabular adjustment. This is the first time these methods have been implemented in a common mathematical and software framework. The software system is referred to as the NCHS confidentiality protection utility for tables (NCHSCPUT).

2. DATA PROTECTION METHODS

This paper describes a software suite comprising five functions for statistical disclosure limitation in a two-dimensional tabular data: suppression; minimum-distance controlled rounding; unbiased controlled rounding; controlled rounding subject to subtotals; and controlled tabular adjustment. The five subsections that follow each describe one of these functions and features of the corresponding software module. Table I presents the original (pre-disclosure limitation) data table comprising an array of 10 rows and 5 columns, plus totals. The cell entries were randomly generated by the software. For convenience, Table I is treated as a table of count

Table I. Original 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
Column sums	3477	5542	5002	5748	6817	26 586

Table II. Results of cell suppression using 5 as the threshold as applied to Table I.

	Col 1	Col 2	Col 3	Col 4	Col 5	Row sums
Row 1	1*	309†	838	369	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
Column sums	3477	5542	5002	5748	6817	26 586

*Primary suppression cell.
†Complementary suppression cell.

data. The five disclosure limitation functions were separately applied to Table I, and Tables II–VI are the resulting disclosure-limited output tables.

Two-dimensional tables enjoy mathematical properties absent from higher-dimensions [3]. Most importantly, a two-dimensional table can be modelled as a *mathematical network* which can result in a considerable reduction in computational time when optimizing certain functions. Complex optimization functions involving integer variables, which could require exponential computing time, can be performed instead by computationally efficient continuous network methods which require only polynomial (quadratic) computing time. It is for this reason, and the fact that two-dimensional tables are ubiquitous in statistical analysis, that the National Center for Health Statistics sponsored the development of this software suite.

Table III. Results of minimum-distance controlled rounding as applied to Table I using base 5 and L_2-norm.

	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
Column sums	3475	5545	5000	5750	6815	26 585

Table IV. Results of unbiased controlled rounding as applied to Table I using base 5 and L_2-norm.

	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
Column sums	3475	5545	5005	5745	6815	26 585

2.1. Complementary cell suppression function

Complementary cell suppression removes from publication the values of all cells representing direct disclosure of confidential data on individual respondents (the *disclosure cells*), together with a sufficient number of appropriately selected non-disclosure cells (the *complementary cells*) 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 determination of which cells are the disclosure cells is made by applying a quantitative *disclosure rule* to the cell data. For count data, typically a non-zero cell is a disclosure cell if its value is below a specified *threshold*, e.g. $n = 5$. Zero cells can be, but most frequently are not, regarded as disclosure cells. For magnitude data, many disclosure rules are possible, but notable is the *p-percent rule* that declares disclosure whenever the cell value minus the contribution of the second largest contributor is less than $(100 + p)$-percent of the contribution of the largest contributor. This rule protects each contributor from having its contribution

Table V. Results of controlled rounding subject to subtotal constraints as applied to Table I using base 5 and L_2-norm.

	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	**10**	**875**	**700**	555	772	2914
Row 7	**950**	**875**	**365**	747	681	3618
Row 8	**130**	**105**	**530**	721	660	2143
Row 9	**140**	**705**	**785**	4	916	2548
Row 10	560	647	633	527	987	3354
Column sums	3477	5542	5002	5748	6817	26 586

Table VI. Results of controlled tabular adjustment (CTA) as applied to Table I.

	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
Column sums	3477	5542	5002	5748	6817	26 586

*Modified cell that was at or below a threshold of 5.
†Other cell that was modified.

divulged to a competitor (for example, a competing hospital in the same geographic area) any closer than p-percent of its value.

If the disclosure rule identifies a cell as a disclosure cell, which is then suppressed, the question arises as to how much additional protection is necessary to reach an acceptable level of disclosure risk. Cox [4] provides a methodology for doing so, which has been incorporated within this software. For the n-threshold rule, the answer is n minus the cell value. For example, if 5 is the minimum number (*threshold*) of respondents required for a published cell and a cell value is 2, then 3 is the least amount of protection that must be added to the cell value. For the p-percent rule, the answer is p-percent of the largest contribution minus the sum of the ‘smaller’ contributions, viz. the 3rd, 4th, etc. largest contributions. The amount of additional protection required for each disclosure cell, called its *protection limit*, must be maintained in each equation (row, column, etc.) containing the cell, as well as in combinations thereof. Because Table I is treated as a table of count data, in the examples to follow we employ a threshold disclosure rule with $n = 5$.

The mathematical and computational challenge of the cell suppression problem is to select complementary cells that provide sufficient disclosure protection while minimizing the amount of information lost due to suppression. Information loss is typically measured as: number of cells suppressed, total value suppressed, total percent of value suppressed, or other functions such as total of logarithm of one plus value suppressed. Typically, an objective function (or cost function) is formed which includes terms related to one of the types of information losses described in the previous sentence. Then, the optimization (cell suppression) is carried out by minimizing the objective function subject to certain constraints, such as row or column totals. The choice of objective function is an important one for the data provider: the mathematical methods and software will produce an optimal solution with equal ease 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 complementary cell suppression method based on mathematical networks of Cox [5] is used as the cell suppression function in the NCHSCPUT. The software performs the optimization via (network) linear programming.

Table II displays the results of applying the *complementary cell suppression* function to Table I. The optimization criterion is to minimize the total value suppressed. The conventional ordered pair notation (row #, column #) will be used to refer to table cells. The counts in cells (1,1), (3,3), and (9,4) are the primary suppressions, and the counts in cells (1,2), (3,2), (8,1), (8,2), (8,3), (8,4), and (9,1) are the complementary suppressions. On a computer screen, the primary suppressions are highlighted in blue, and the complementary suppressions are highlighted in red. Using the optimization criterion to minimize the total value suppressed as was done above, the total complementary value suppressed is 2093. Had the provider specified instead the objective function 'minimum number of cells suppressed', complementary suppressions would have been made at either (1, 3), (3, 4) and (9, 1) or at (1, 4), (3, 1) and (9, 3) with total complementary values suppressed equal to 1778 and 1499, respectively. [*Note*: Cell suppression assigns a cost to each cell according to that cell's impact on the sums. The algorithm then uses a network flow procedure to find the flow with the lowest cost to protect one of the primary suppressed cells. This network is used iteratively on each of the cells requiring protection (primary suppressions), each time using the information from the previous iteration and assigning cells that have been suppressed a negligible cost so that they are used first. This algorithm may not result in the smallest overall sum, due to its iterative nature. Iteration is necessary because the problem of simultaneous overall protection has been shown to be NP-hard.]

2.2. Controlled rounding function

Controlled rounding replaces each entry (including totals) in a one- or two-way tabular array A by an integer multiple of a specified positive integer *rounding base* B subject to the following requirements:

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

(b) The sum of the rounded values for any row (or column) of A equals the rounded value of the corresponding row (or column) total entry. 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, $1 \leqslant p < \infty$, 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 table A. That is, the objective function to minimize with respect to the L_p-norm is

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

where $R(A)$ represents the rounded table for table A and $R(a_{ij})$ denotes the rounded cell entry for cell entry a_{ij}.

Cox and Ernst [6] 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 statistical, interpretation (namely, minimum of a corresponding Euclidean distance). However, other linear functions can be employed to optimize statistical properties, (e.g. sum of differences between rounded and unrounded entries minimizes change to the table mean resulting from rounding).

The *controlled rounding* function that is used in the software is based on the methodology described by Cox and Ernst [6] and Causey *et al.* [7]. Unfortunately, controlled rounding cannot be extended to three- or higher-dimensional tables in all cases [8]. Often, exact methods for two-dimensional tables as described here are applied iteratively to the vertical planes of a three-dimensional table to produce approximate results. Table III displays the results of applying the *controlled rounding* function (to base $B = 5$ with power $p = 2$) to Table I.

The following computation times were computed using a Pentium 4 processor with 261 200 KB of RAM:

- 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 min.
- A table with 1000 rows and 5 columns was rounded in 1 h and 40 min.

2.3. Unbiased controlled rounding function

Given a two-dimensional table A, the objective of *unbiased controlled rounding* is to construct a controlled rounding table $R(A)$ of A satisfying: $E[R(a_{ij})] = a_{ij}$ for all table entries. In other words, in lieu of optimizing with respect to a Euclidean measure of distance, the desired solution preserves original values with respect to the statistical criterion *expectation*. This method relies on a probability measure on rounding down or up for each table entry, as follows. Let r denote the remainder of the cell value after division by the rounding base B; $0 \leqslant r < B$. Round the cell down with probability $(b - r)/B$; round the cell up with probability r/B. The implementation does not rely on linear programming *per se* but on a specialized algorithm from mathematical programming known as the *stepping stones algorithm*. See Reference [9] for details.

In summary, the conditions for unbiased controlled rounding are that every entry a_{ij} of table A satisfies the following:

(a) $R(a_{ij}) = B[a_{ij}/B]$ or $B([a_{ij}/B] + 1)$,
(b) $R(A)$ is additive,
(c) $|R(a_{ij}) - a_{ij}| < B$,
(d) $E[R(a_{ij})] = a_{ij}$.

The unbiased controlled rounding function that is used in the software is based on the methodology of Cox [9]. Table IV displays the results of applying the *unbiased controlled rounding* function to Table I (to base $B = 5$ with power $p = 2$).

Here are some computational times for the unbiased controlled rounding function

- A table with 50 rows and 50 columns was rounded in 1 s.
- A table with 100 rows and 100 columns was rounded in 4 s.
- A table with 400 rows and 25 columns was rounded in 5 s.
- A table with 2000 rows and 25 columns was rounded in 5 min and 45 s.

2.4. Controlled rounding subject to subtotal constraints

The *controlled rounding subject to subtotal constraints* function that is used in the software is based on the methodology described by Cox and George [10]. The methodology used in this function is similar to that used for *controlled rounding* as discussed earlier, but seeks as well to preserve additivity to subtotals along rows or columns of the table, at least to within 'base' units of the original subtotal. Cox and George [10] show how this can be done along any number of rows or columns of the table, and in addition that one may fail if trying to preserve subtotals along both rows and columns. Recall that controlled rounding for a two-way table was presented as a capacitated transportation problem. This function extends that methodology to tables with subtotals along one, but not both, dimensions.

Table V displays the results of applying the *controlled rounding subject to subtotal constraints* (to base $B = 5$ with power $p = 2$) function to the first three cell entries in each of rows 6–9 of Table I in order to preserve the subtotals across the first three columns for each of the selected rows. The original subtotals in Table I for the first three entries in rows 6–9 are 1587, 2190, 762, and 1628, respectively. The new corresponding subtotals in Table V are 1585, 2190, 765, and 1630, respectively, so that each subtotal was kept within base $B = 5$ of the original.

2.5. Controlled tabular adjustment function

A new disclosure limitation method, *controlled tabular adjustment* (CTA), was introduced by Dandekar and Cox [11] as an alternative to complementary cell suppression. The Dandekar–Cox method 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. Adjustments to other values are controlled by capacitating changes to be small. In the output file (table) on the computer screen, the disclosure cells are highlighted in blue while the other cells that are modified are highlighted in red.

In the current software implementation, a factor is used to randomize the results of tabular adjustment. Therefore, performing the procedure more than once on the same input file will generate different outputs. Table VI shows the results of applying CTA to Table I.

3. SOFTWARE FEATURES

As mentioned in the Introduction of this paper, two-dimensional statistical tables are a staple of official statistics. The NCHS has sponsored the development of disclosure limitation software for two-dimensional tables by OptTek Systems, Inc. This section will briefly describe some of the main features of the software and its different functions: cell suppression; minimum-distance controlled rounding; unbiased controlled rounding; controlled rounding subject to subtotal constraints; and controlled tabular adjustment. This is the first time these methods have been implemented in a common mathematical and software framework. The software system is referred to as the NCHSCPUT. Figure 1 provides an image of the NCHSCPUT

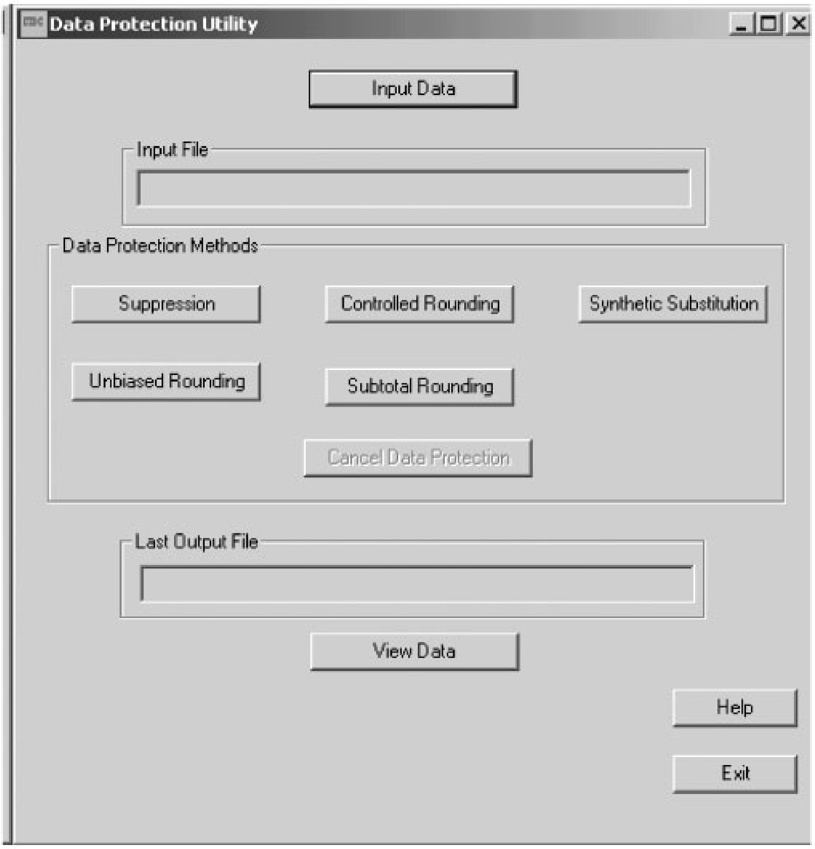

Figure 1. NCHSCPUT screen.

screen. The following is a brief summary of some of the software features. Additional details are available using the '*Help*' button.

The '*Input Data*' button allows the user to select or create data using the Data Editor. The '*Data Protection Methods*' button allows the user to select the type of protection the user would like to use. The output can then be saved in different files if the user wants to compare the results. The '*View Data*' button allows the user to display the output data created from the data protection method used and allows the user to view the output from past runs with the aid of the NCHSCPUT Data Editor. The Data Editor also permits the user to open a file that has been previously created by the Data Editor, generate random data, insert data, modify data, select the base to use in all of the data protection methods, and select the power to use for all of the controlled rounding methods. The Data Editor can be resized so that more than one dialog can fit on the screen or more data can be seen without scrolling. After the data have been entered or modified, they may be saved in the current file using the '*Save*' button, or in a different file using the '*Save As*' button.

Once the data have been saved to a file, they can be exported into a file that can be read by various software packages. The export file is tab-delineated with one line per row and the columns separated by tabs. This allows the user to then open the file in applications such as NotePad or Excel and peruse the data in a familiar format, or print the file using that application's printing capabilities. For suppressed cells, the primary suppressions are prefaced with 'PS' and the complementary suppressions are prefaced with 'CS'.

The NCHSCPUT also imports files created by other programs such as SAS. The input file's format must have a line for each data row. Columns in a row are separated by blanks. The input file should not include row and column sums.

The NCHSCPUT software is available free of charge upon request. However, no technical support is available from the NCHS or OptTek Systems, Inc. For further information, contact the lead author.

4. FUTURE RESEARCH AND DEVELOPMENT

The software developed for this project is a tool which features some of the different mathematical functions for protecting potential disclosure cell values in two-way tables. The ultimate goal of this project is to develop production level software that can be embedded into NCHS data analysis activities, for example, the NCHS Research Data Center (RDC). Disclosure limitation research, controlled tabular adjustment in particular, is ongoing, and is addressing both data quality and data confidentiality concerns, e.g. [12]. We expect to incorporate new and improved methods into the software as they become available.

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. Cox LH. On properties of multi-dimensional statistical tables. *Journal of Statistical Planning and Inference* 2003; **117**(2):251–273.
4. Cox LH. Linear sensitivity measures in statistical disclosure control. *Journal of Statistical Planning and Inference* 1981; **5**:153–164.
5. Cox LH. Network models for complementary cell suppression. *Journal of the American Statistical Association* 1995; **90**:1453–1462.

6. Cox LH, Ernst LR. Controlled rounding. *INFOR* 1982; **20**:423–432.
7. Causey BD, Cox LH, Ernst LR. Applications of transportation theory to statistical problems. *Journal of the American Statistical Association* 1985; **80**:903–909.
8. Ernst LR. Further applications of linear programming to sampling problems. *Technical Report-Census/SRD/RR*-89-05. U.S. Census Bureau, Department of Commerce, Washington, DC, 1989. Available: http://www.census.gov/srd/www/byname.html
9. Cox LH. A constructive procedure for unbiased controlled rounding. *Journal of the American Statistical Association* 1987; **82**:520–524.
10. Cox LH, George JA. Controlled rounding for tables with subtotals. *Annals of Operations Research* 1989; **20**:141–157.
11. Dandekar RA, Cox LH. Synthetic tabular data—an alternative to complementary cell suppression. 27 February, 2002, unpublished manuscript.
12. 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), Luxembourg, April 2003, to appear. Available: http://www.unece.org/stats/documents/2003.04.confidentiality.htm