Protecting Sensitive Tabular Data by Complementary Cell Suppression - Myth & Reality

Ramesh A. Dandekar

Energy Information Administration, 1000 Independence Avenue, Washington DC 20585
Ramesh.Dandekar@EIA.DOE.gov (http://mysite.verizon.net/vze7w8vk/)

INTRODUCTION

Title 13, U.S.C., Section 9 and the newly adopted CIPSEA of 2002 impose heavy financial fines and prison penalties for a public disclosure of sensitive statistical information. Currently, complementary cell suppression procedures are mostly used by statistical agencies to protect sensitive tabular data from disclosure. It is generally believed that the linear programming (LP) based complementary cell suppression procedures offer the best protection from wrongful disclosure of statistical information. In recent years LP-based automated audit procedures have been advocated and are being used to ensure the adequacy of protection offered by cell suppression patterns. LP-based lower and upper bounds for suppressed tabular cells are typically used to determine the adequacy of disclosure control measures. This paper identifies limitations of conclusions drawn using LP-based audit procedures. We utilize commonly used analytical procedures to demonstrate the relative ease with which statistical disclosure of sensitive tabular data could occur. We conclude by providing additional safeguard measures required to avoid such disclosures.

CURRENT PRACTICE

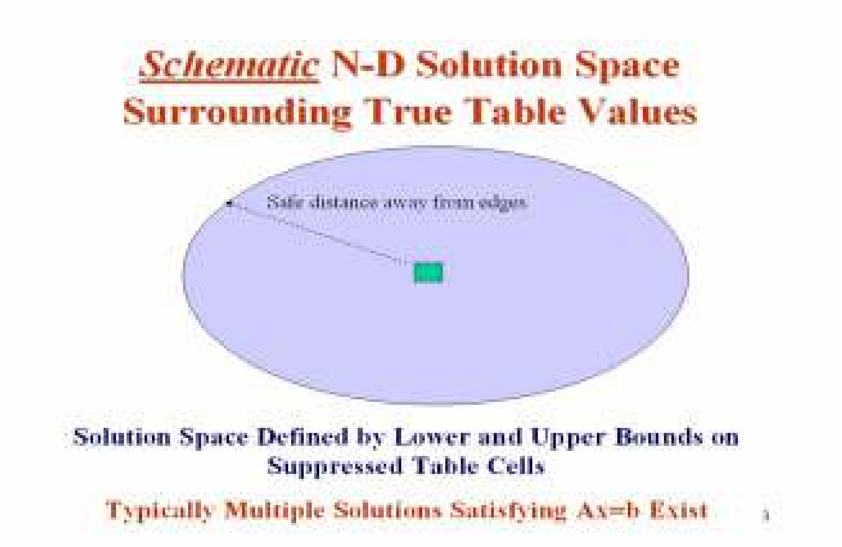
The complementary cell suppression methods, as currently practiced by national statistical offices (NSO), enable data users to determine a multi-dimensional solution space surrounding the "incomplete" tabulation available in the public domain. Linear programming (LP) based lower and upper bounds on the withheld tabular cells are used to establish the boundaries for the solution space.

NSOs are required to ensure that the real complete table containing sensitive cells is well hidden inside the solution space a safe distance away from the edges of the solution space. The solution space typically contains multiple feasible solutions that satisfy the equality constraints associated with the complete real table structure.

Feasible solutions residing close to the edges of the solution space tend to yield poor estimates of the values of withheld cells. On the other hand, feasible solutions located away from the edges of the solution space and toward the "centroid" of the solution space tend to be of better quality and more closely resemble the hidden real complete table. ***This phenomenon has the potential to cause the disclosure of sensitive tabular data protected by complementary cell suppression methods***.

Typically in an attempt to minimize the information loss, NSOs are under pressure to avoid over protection of sensitive tabular cells. The over protection of sensitive tabular cells results in an increase in the size of the solution space.

As per current practice, the solution space is expected to be "just right" in size. Smaller than a minimum required solution space, determined by LP-based lower and upper bounds, is known to be unacceptable. Larger than a minimum required solution space, determined by LP-based lower and upper bounds, is thought to cause unnecessary information loss. As a result, in recent years much of the efforts in tabular data protection area have been concentrated in keeping the cell suppression related solution space to a bare minimum.



Schematic N-D Solution Space Surrounding True Table Values

Solution Space Defined by Lower and Upper Bounds on Suppressed Table Cells

Typically Multiple Solutions Satisfying Ax=b Exist

CURRENT TOOLS

Optimization Technology Center of Northwestern University and Argonne National Laboratory at http://www-unix.mcs.anl.gov/otc/Guide/faq/ describes linear programming tools as follows:

"Two families of solution techniques are in wide use today. Both visit a progressively improving series of trial solutions, until a solution is reached that satisfies the conditions for an optimum. ***Simplex methods***, introduced by Dantzig about 50 years ago, visit "basic" solutions computed by fixing enough of the variables at their bounds to reduce the constraints Ax = b to a square system, which can be solved for unique values of the remaining variables. Basic solutions represent extreme boundary points of the feasible region defined by Ax = b, x >= 0, and the simplex method can be viewed as moving from one such point to another along the edges of the boundary. ***Barrier or interior-point methods, by contrast, visit points within the interior of the feasible region. …….***"

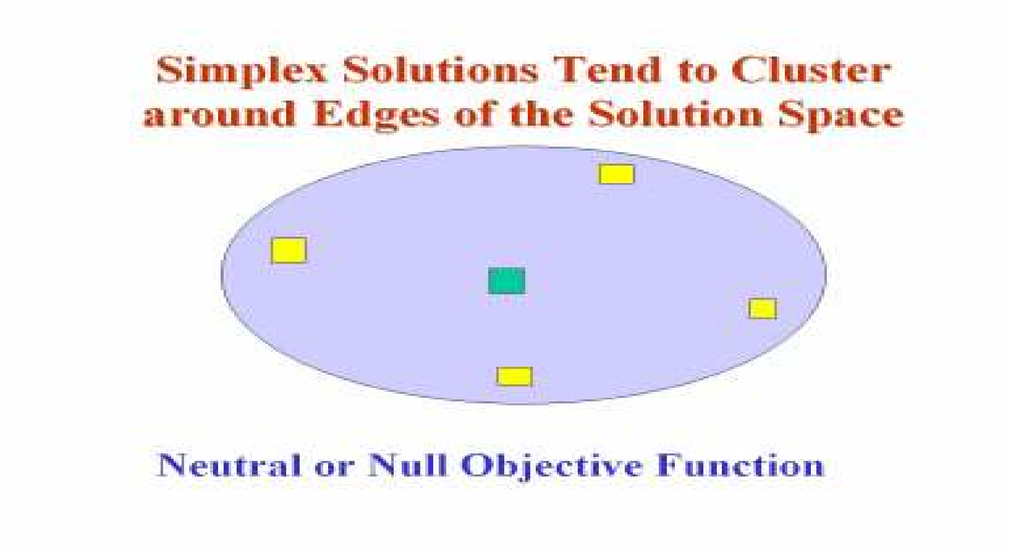
The increased potential for statistical disclosure of the withheld sensitive tabular data is directly related to the basic property of interior-point methods to ***visit points within the interior of the feasible region,*** where the real complete table containing sensitive tabular cells resides.

We use the following simple illustrative example supplied by Prof. Jordi Castro http://www-eio.upc.es/~jcastro/ to further clarify the difference in the working of two families of LP solvers.

min 0
st. $x1 + x2 + x3 = 3$
$x1, x2, x3 > = 0$

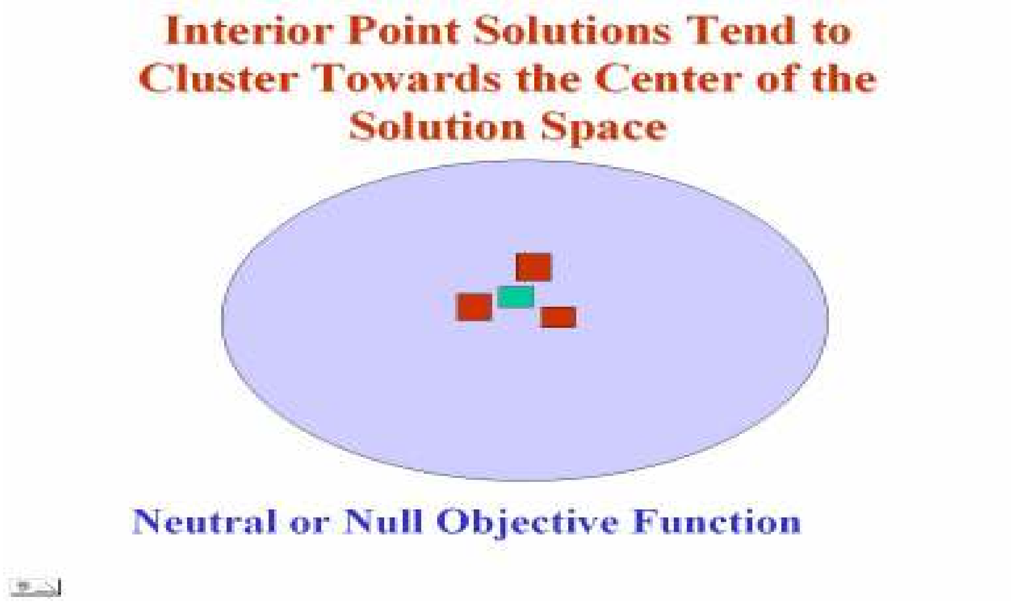
Interior point methods will provide the solution $x1 = x2 = x3 = 1$

The simplex methods will provide some $xi = 3$, the other two $xj = 0$.



Simplex Solutions Tend to Cluster around Edges of the Solution Space

Neutral or Null Objective Function



Interior Point Solutions Tend to Cluster Towards the Center of the Solution Space

Neutral or Null Objective Function

A knowledgeable individual can easily exploit the working knowledge of interior-point methods to obtain “high quality” additive point estimates for missing tabular cells by (1) not specifying the objective function (or by using a dummy objective function) and (2) capturing the first feasible solution that satisfies the tabular data equality constraints. A moderately sized solution space, in combination with the tendency of interior point methods to the visit interior of the feasible region, will always ensure high precision estimates. These estimates are most likely to cause the statistical disclosure of withheld sensitive cells.

ILLUSTRATIVE EXAMPLE

In Table 1 we have used the 3-D tabular data example from Dandekar/Cox (2002) paper available from http://mysite.verizon.net/vze7w8vk/ to illustrate the severity of the disclosure problem associated with current SDL practice. The table contains 24 sensitive cells. The table is protected by using 44 complementary cell suppressions. Table 2 shows the LP-based lower and upper bounds for the 24 sensitive cells. The p percent rule (p=10%) was used to identify the sensitive cells. Except for two minor violations for sensitive cell #6 and #18, the suppression pattern associated with the 44 complementary cells fully satisfies the current requirement for "safe table".

STATISTICAL ESTIMATION

Typically, statistical estimates for missing table cell values can be derived by using 1) additive point estimates 2) method of averages and 3) peak densities associated with frequency distributions. The last two methods, by themselves, do not provide additive tabular estimates. However, when combined with the controlled tabular adjustment (CTA) method of Dandekar/Cox (2002), the last two methods are capable of providing additive tabular estimates.

We have used the interior-point based, PCx linear programming solver available from http://www-fp.mcs.anl.gov/otc/Tools/PCx/ to illustrate the severity of the disclosure problem resulting from statistical estimates for sensitive table cells.

Table 3 provides additive point estimates for missing sensitive cells[1] by using the conventional simplex method and the PCx solver. The null-objective function was used to derive the additive point estimates. Three of the simplex estimates and 14 of the PCx estimates violate protection level for the sensitive cell causing statistical disclosure. These findings are consistent with the properties associated with the two families of solution techniques as described on the Argonne National Laboratory web site above.

Table 4 provides statistics based on averages from 138 LP solutions obtained by using the PCx software. Half of the LP solutions (sixty-nine) were for a minimization of the objective function. The remaining LP solutions were for a maximization of the objective function. Sixty-eight solutions in each group were obtained by using only one variable in the objective function. One solution in each group included all the sixty-eight variables in the objective function. Sixteen of the twenty-four averages are within the prohibited protection range causing the statistical disclosure of 16 sensitive cells.

Table 5 uses the outcome from the same 138 LP solutions to generate the frequency distribution of estimates for missing sensitive cells. The table contains three lines of output for every sensitive cell. The first line in the table displays the true cell value of the sensitive cell (714 for the first sensitive cell) and the LP-based audit range (409 for the first sensitive cell).

In the next two lines we divide the audit range into ten equal intervals and summarize the frequency count resulting from the 138 LP runs. The first line shows the actual count, while the second line shows the interval values associated with the count. For the first sensitive cell, the peak density of 97 is within the sixth interval ranging from 697 to 738. The comparison of the location of the peak of the density function relative to the true cell value reveals statistical disclosure for almost all of the twenty-four sensitive cells.

TARGETING THE CENTROID OF THE SOLUTION SPACE

Knowing that the real complete table is typically hidden some where in the vicinity of the centroid of the solution space, a knowledgeable individual can also use any general purpose LP solver (not necessarily interior point solver) to derive "high precision" additive point estimates for the suppressed tabular cells. Related mathematical formulation requires that each suppressed tabular cell (X_{estimate}) be represented by three variables in the tabular data equality constraints, namely X_{centroid}, Y_{plus} and Y_{minus}.

[1] Space limitations prohibit us from providing values for non-sensitive tabular cells.

Where $X_{centroid} = 0.5 * X_{lower\_LP\_bound} + 0.5 * X_{upper\_LP\_bound}$,

$X_{estimate} = X_{centroid} + Y_{plus} - Y_{minus}$ and

Y_{plus} and Y_{minus} are minimal plus or minus corrective adjustments required to ensure additivity of tabular cells

An individual with advanced computation skills could even go further and use either random Monte Carlo simulations or some sophisticated stratification scheme to obtain density functions (and peak density values) for the missing table cell values by using the following simple equation:

$X_{centroid} = R * X_{lower\_LP\_bound} + (1.0 – R) * X_{upper\_LP\_bound}$

Where R = Random Number between zero and one

If the individual further decides to restrict the search for the feasible solution, say to within a 10 percentile range around the centroid of the solution space, then the values for the random number could be restricted to within 0.4 and 0.6 to achieve that objective.

CONCLUSIONS AND RECOMMENDATIONS

As a result of the easy access to the interior-point methods, such as PCx software tool, the LP-based lower and upper bounds of tabular data cell suppression patterns can no longer be used *alone* to judge the adequacy of the cell suppression pattern.

Conventional statistical analytical measures such as additive point estimates, method of averages and peak density values associated with frequency distributions, in combination with interior point methods, could be used with trivial efforts to cause a statistical disclosure of sensitive tabular data.

Contrary to current belief, over protection of the sensitive tabular data ***reduces*** the possibility of statistical disclosure resulting from use of interior point LP solvers. As a result, the over protection of sensitive tabular data is no longer an undesirable property of cell suppression pattern.

The current practice of using relatively small size cells as complementary suppression cells has a ***tendency*** to produce tighter LP bounds with sharp peak density functions. Therefore, this practice should be used with caution.

Use of cost functions such as reciprocal of cell value or log(cell value)/cell value to develop complementary cell suppression pattern targets large size cells. Complementary cell suppression pattern based on these functions has a ***tendency*** to produce wider protection intervals with flatter density functions. For this reason, these cost functions should be given a serious consideration.

With new technical challenges arising from the easy access to interior point methods, NSOs might want to explore the possibility of switching form the complementary cell suppression methods to other tabular data protection methods.

Emerging methods such as synthetic tabular data, which also is referred to as controlled tabular adjustment (CTA), offers sensitive tabular data required protection from disclosure without disclosing the solution space associated with the CTA pattern. The lack of complete information pertaining to the solution space associated with CTA pattern eliminates the possibility of the outside user deploying standardized external procedures to estimate true value for sensitive cells on a massive scale.

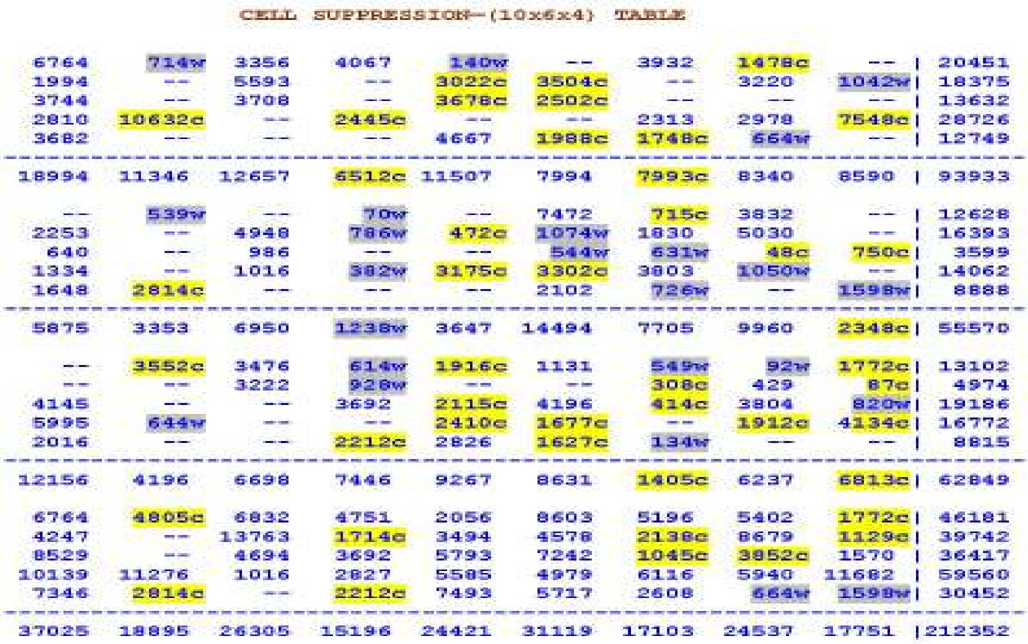
REFERENCES

Dandekar R. A. and Cox L. H. (2002), Synthetic Tabular Data: An Alternative to Complementary Cell Suppression, manuscript available from ramesh.dandekar@eia.doe.gov or from URL http://mysite.verizon.net/vze7w8vk/

Dandekar, R.A (2003), Cost Effective Implementation of Synthetic Tabulation (a.k.a. Controlled Tabular Adjustments) in Legacy and New Statistical Data Publication Systems, working paper 40, UNECE Work session on statistical data confidentiality (Luxembourg, 7-9 April 2003) http://epp.eurostat.cec.eu.int/portal/page?_pageid=1073,1135281,1073_1135295&_dad=portal&_schema=PORTAL&p_product_code=KS-CR-03-004-3

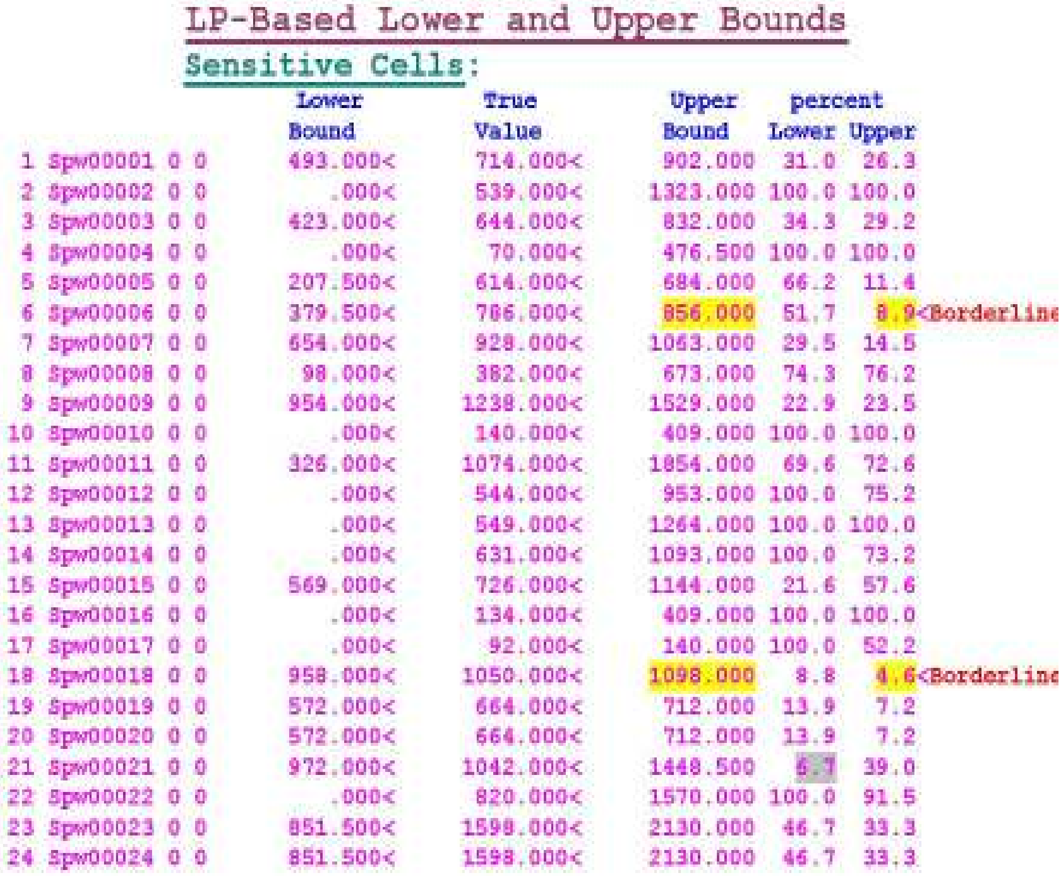
Dandekar Ramesh A. (2004), Maximum Utility-Minimum Information Loss Table Server Design for Statistical Disclosure Control of Tabular Data, pp 121-135, Lecture Notes in Computer Science, Publisher: Springer-Verlag Heidelberg, ISSN: 0302-9743, Volume 3050 / 2004, Title: Privacy in Statistical Databases: CASC Project International Workshop, PSD 2004, Barcelona, Spain, June 9-11, 2004.

Table 1



```
                        CELL  SUPPRESSION—(10x6x4)  TABLE

   6764     714w    3356     4067     140w      --     3932    1478c      -- |  20451
   1994      --     5593      --     3022c    3504c      --     3220    1042w|  18375
   3744      --     3708      --     3678c    2502c      --      --       -- |  13632
   2810   10632c     --     2445c      --       --     2313     2978    7548c|  28726
   3682      --      --       --     4667     1988c    1748c    664w      -- |  12749
--------------------------------------------------------------------------------------
  18994    11346   12657    6512c   11507     7994    7993c    8340     8590 |  93933

     --     539w      --      70w      --     7472     715c    3832      -- |  12628
   2253      --     4948     786w     472c    1074w    1830     5030      -- |  16393
    640      --      986      --       --      544w    631w      48c    750c|   3599
   1334      --     1016     382w    3175c    3302c    3803    1050w      -- |  14062
   1648    2814c      --      --       --     2102     726w      --    1598w|   8888
--------------------------------------------------------------------------------------
   5875     3353    6950    1238w    3647    14494     7705     9960    2348c|  55570

     --    3552c    3476     614w    1916c    1131     549w      92w   1772c|  13102
     --      --     3222     928w      --       --     308c     429      87c|   4974
   4145      --       --     3692    2115c    4196     414c    3804     820w|  19186
   5995     644w      --       --    2410c    1677c      --    1912c   4134c|  16772
   2016      --       --    2212c    2826     1627c    134w      --       -- |   8815
--------------------------------------------------------------------------------------
  12156     4196    6698     7446    9267     8631    1405c    6237    6813c|  62849

   6764    4805c    6832     4751    2056     8603     5196     5402   1772c|  46181
   4247      --    13763    1714c    3494     4578    2138c    8679    1129c|  39742
   8529      --     4694     3692    5793     7242    1045c    3852c    1570 |  36417
  10139    11276    1016     2827    5585     4979     6116     5940   11682 |  59560
   7346    2814c      --    2212c    7493     5717     2608     664w   1598w|  30452
--------------------------------------------------------------------------------------
  37025    18895   26305    15196   24421    31119    17103    24537   17751 |212352
```

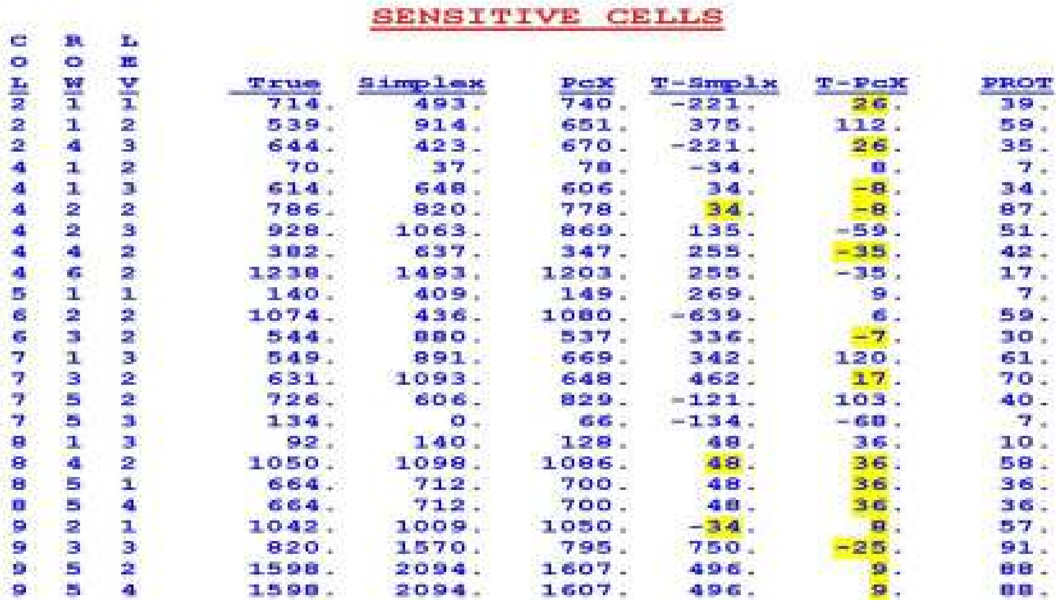
Table 2



```
                    LP-Based Lower and Upper Bounds
                    Sensitive Cells:
                           Lower           True             Upper    percent
                           Bound           Value            Bound   Lower Upper
  1 Spw00001 0 0         493.000<        714.000<         902.000  31.0  26.3
  2 Spw00002 0 0            .000<        539.000<        1323.000 100.0 100.0
  3 Spw00003 0 0         423.000<        644.000<         832.000  34.3  29.2
  4 Spw00004 0 0            .000<         70.000<         476.500 100.0 100.0
  5 Spw00005 0 0         207.500<        614.000<         684.000  66.2  11.4
  6 Spw00006 0 0         379.500<        786.000<         856.000  51.7   8.9<Borderline
  7 Spw00007 0 0         654.000<        928.000<        1063.000  29.5  14.5
  8 Spw00008 0 0          98.000<        382.000<         673.000  74.3  76.2
  9 Spw00009 0 0         954.000<       1238.000<        1529.000  22.9  23.5
 10 Spw00010 0 0            .000<        140.000<         409.000 100.0 100.0
 11 Spw00011 0 0         326.000<       1074.000<        1854.000  69.6  72.6
 12 Spw00012 0 0            .000<        544.000<         953.000 100.0  75.2
 13 Spw00013 0 0            .000<        549.000<        1264.000 100.0 100.0
 14 Spw00014 0 0            .000<        631.000<        1093.000 100.0  73.2
 15 Spw00015 0 0         569.000<        726.000<        1144.000  21.6  57.6
 16 Spw00016 0 0            .000<        134.000<         409.000 100.0 100.0
 17 Spw00017 0 0            .000<         92.000<         140.000 100.0  52.2
 18 Spw00018 0 0         958.000<       1050.000<        1098.000   8.8   4.6<Borderline
 19 Spw00019 0 0         572.000<        664.000<         712.000  13.9   7.2
 20 Spw00020 0 0         572.000<        664.000<         712.000  13.9   7.2
 21 Spw00021 0 0         972.000<       1042.000<        1448.500   6.7  39.0
 22 Spw00022 0 0            .000<        820.000<        1570.000 100.0  91.5
 23 Spw00023 0 0         851.500<       1598.000<        2130.000  46.7  33.3
 24 Spw00024 0 0         851.500<       1598.000<        2130.000  46.7  33.3
```

Table 3

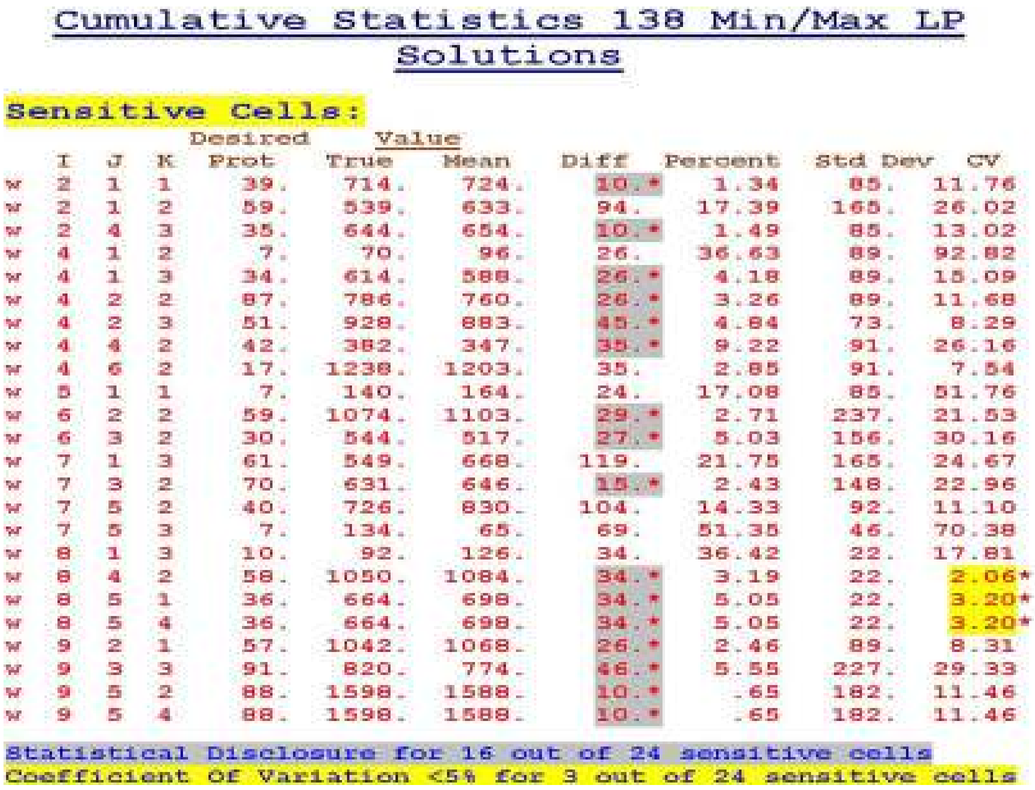
Additive Estimates Simplex versus Interior Point Method



SENSITIVE CELLS

COL	ROW	LEV	True	Simplex	PcX	T-Smplx	T-PcX	PROT
2	1	1	714.	493.	740.	-221.	26.	39.
2	1	2	539.	914.	651.	375.	112.	59.
2	4	3	644.	423.	670.	-221.	26.	35.
4	1	2	70.	37.	78.	-34.	8.	7.
4	1	3	614.	648.	606.	34.	-8.	34.
4	2	2	786.	820.	778.	34.	-8.	87.
4	2	3	928.	1063.	869.	135.	-59.	51.
4	4	2	382.	637.	347.	255.	-35.	42.
4	6	2	1238.	1493.	1203.	255.	-35.	17.
5	1	1	140.	409.	149.	269.	9.	7.
6	2	2	1074.	436.	1080.	-639.	6.	59.
6	3	2	544.	880.	537.	336.	-7.	30.
7	1	3	549.	891.	669.	342.	120.	61.
7	3	2	631.	1093.	648.	462.	17.	70.
7	5	2	726.	606.	829.	-121.	103.	40.
7	5	3	134.	0.	66.	-134.	-68.	7.
8	1	3	92.	140.	128.	48.	36.	10.
8	4	2	1050.	1098.	1086.	48.	36.	58.
8	5	1	664.	712.	700.	48.	36.	36.
8	5	4	664.	712.	700.	48.	36.	36.
9	2	1	1042.	1009.	1050.	-34.	8.	57.
9	3	3	820.	1570.	795.	750.	-25.	91.
9	5	2	1598.	2094.	1607.	496.	9.	88.
9	5	4	1598.	2094.	1607.	496.	9.	88.

Table 4



Cumulative Statistics 138 Min/Max LP Solutions

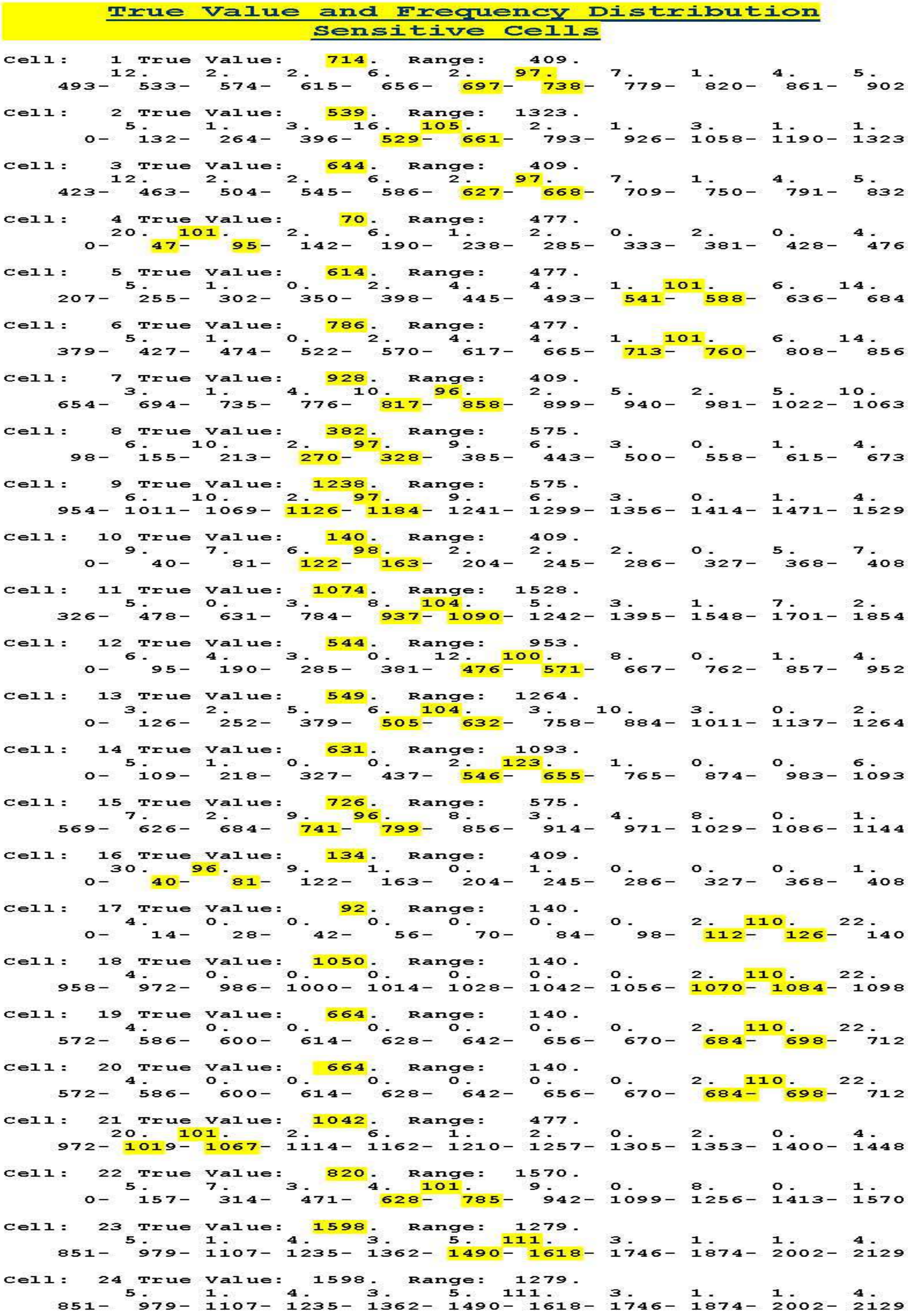
Sensitive Cells:

				Desired	Value					
	I	J	K	Prot	True	Mean	Diff	Percent	Std Dev	CV
w	2	1	1	39.	714.	724.	10.*	1.34	85.	11.76
w	2	1	2	59.	539.	633.	94.	17.39	165.	26.02
w	2	4	3	35.	644.	654.	10.*	1.49	85.	13.02
w	4	1	2	7.	70.	96.	26.	36.63	89.	92.82
w	4	1	3	34.	614.	588.	26.*	4.18	89.	15.09
w	4	2	2	87.	786.	760.	26.*	3.26	89.	11.68
w	4	2	3	51.	928.	883.	45.*	4.84	73.	8.29
w	4	4	2	42.	382.	347.	35.*	9.22	91.	26.16
w	4	6	2	17.	1238.	1203.	35.	2.85	91.	7.54
w	5	1	1	7.	140.	164.	24.	17.08	85.	51.76
w	6	2	2	59.	1074.	1103.	29.*	2.71	237.	21.53
w	6	3	2	30.	544.	517.	27.*	5.03	156.	30.16
w	7	1	3	61.	549.	668.	119.	21.75	165.	24.67
w	7	3	2	70.	631.	646.	15.*	2.43	148.	22.96
w	7	5	2	40.	726.	830.	104.	14.33	92.	11.10
w	7	5	3	7.	134.	65.	69.	51.35	46.	70.38
w	8	1	3	10.	92.	126.	34.	36.42	22.	17.81
w	8	4	2	58.	1050.	1084.	34.*	3.19	22.	2.06*
w	8	5	1	36.	664.	698.	34.*	5.05	22.	3.20*
w	8	5	4	36.	664.	698.	34.*	5.05	22.	3.20*
w	9	2	1	57.	1042.	1068.	26.*	2.46	89.	8.31
w	9	3	3	91.	820.	774.	46.*	5.55	227.	29.33
w	9	5	2	88.	1598.	1588.	10.*	.65	182.	11.46
w	9	5	4	88.	1598.	1588.	10.*	.65	182.	11.46

Statistical Disclosure for 16 out of 24 sensitive cells

Coefficient Of Variation <5% for 3 out of 24 sensitive cells

Table 5



```
        True Value and Frequency Distribution
                       Sensitive Cells
Cell:    1 True Value:    714.   Range:    409.
        12.    2.    2.    6.    2.   97.    7.    1.    4.    5.
    493-  533-  574-  615-  656-  697-  738-  779-  820-  861-   902

Cell:    2 True Value:    539.   Range:   1323.
         5.    1.    3.   16.  105.    2.    1.    3.    1.    1.
      0-  132-  264-  396-  529-  661-  793-  926- 1058- 1190-  1323

Cell:    3 True Value:    644.   Range:    409.
        12.    2.    2.    6.    2.   97.    7.    1.    4.    5.
    423-  463-  504-  545-  586-  627-  668-  709-  750-  791-   832

Cell:    4 True Value:     70.   Range:    477.
        20.  101.    2.    6.    1.    2.    0.    2.    0.    4.
      0-   47-   95-  142-  190-  238-  285-  333-  381-  428-   476

Cell:    5 True Value:    614.   Range:    477.
         5.    1.    0.    2.    4.    4.    1.  101.    6.   14.
    207-  255-  302-  350-  398-  445-  493-  541-  588-  636-   684

Cell:    6 True Value:    786.   Range:    477.
         5.    1.    0.    2.    4.    4.    1.  101.    6.   14.
    379-  427-  474-  522-  570-  617-  665-  713-  760-  808-   856

Cell:    7 True Value:    928.   Range:    409.
         3.    1.    4.   10.   96.    2.    5.    2.    5.   10.
    654-  694-  735-  776-  817-  858-  899-  940-  981- 1022-  1063

Cell:    8 True Value:    382.   Range:    575.
         6.   10.    2.   97.    9.    6.    3.    0.    1.    4.
     98-  155-  213-  270-  328-  385-  443-  500-  558-  615-   673

Cell:    9 True Value:   1238.   Range:    575.
         6.   10.    2.   97.    9.    6.    3.    0.    1.    4.
    954- 1011- 1069- 1126- 1184- 1241- 1299- 1356- 1414- 1471-  1529

Cell:   10 True Value:    140.   Range:    409.
         9.    7.    6.   98.    2.    2.    2.    0.    5.    7.
      0-   40-   81-  122-  163-  204-  245-  286-  327-  368-   408

Cell:   11 True Value:   1074.   Range:   1528.
         5.    0.    3.    8.  104.    5.    3.    1.    7.    2.
    326-  478-  631-  784-  937- 1090- 1242- 1395- 1548- 1701-  1854

Cell:   12 True Value:    544.   Range:    953.
         6.    4.    3.    0.   12.  100.    8.    0.    1.    4.
      0-   95-  190-  285-  381-  476-  571-  667-  762-  857-   952

Cell:   13 True Value:    549.   Range:   1264.
         3.    2.    5.    6.  104.    3.   10.    3.    0.    2.
      0-  126-  252-  379-  505-  632-  758-  884- 1011- 1137-  1264

Cell:   14 True Value:    631.   Range:   1093.
         5.    1.    0.    0.    2.  123.    1.    0.    0.    6.
      0-  109-  218-  327-  437-  546-  655-  765-  874-  983-  1093

Cell:   15 True Value:    726.   Range:    575.
         7.    2.    9.   96.    8.    3.    4.    8.    0.    1.
    569-  626-  684-  741-  799-  856-  914-  971- 1029- 1086-  1144

Cell:   16 True Value:    134.   Range:    409.
        30.   96.    9.    1.    0.    1.    0.    0.    0.    1.
      0-   40-   81-  122-  163-  204-  245-  286-  327-  368-   408

Cell:   17 True Value:     92.   Range:    140.
         4.    0.    0.    0.    0.    0.    0.    2.  110.   22.
      0-   14-   28-   42-   56-   70-   84-   98-  112-  126-   140

Cell:   18 True Value:   1050.   Range:    140.
         4.    0.    0.    0.    0.    0.    0.    2.  110.   22.
    958-  972-  986- 1000- 1014- 1028- 1042- 1056- 1070- 1084-  1098

Cell:   19 True Value:    664.   Range:    140.
         4.    0.    0.    0.    0.    0.    0.    2.  110.   22.
    572-  586-  600-  614-  628-  642-  656-  670-  684-  698-   712

Cell:   20 True Value:    664.   Range:    140.
         4.    0.    0.    0.    0.    0.    0.    2.  110.   22.
    572-  586-  600-  614-  628-  642-  656-  670-  684-  698-   712

Cell:   21 True Value:   1042.   Range:    477.
        20.  101.    2.    6.    1.    2.    0.    2.    0.    4.
    972- 1019- 1067- 1114- 1162- 1210- 1257- 1305- 1353- 1400-  1448

Cell:   22 True Value:    820.   Range:   1570.
         5.    7.    3.    4.  101.    9.    0.    8.    0.    1.
      0-  157-  314-  471-  628-  785-  942- 1099- 1256- 1413-  1570

Cell:   23 True Value:   1598.   Range:   1279.
         5.    1.    4.    3.    5.  111.    3.    1.    1.    4.
    851-  979- 1107- 1235- 1362- 1490- 1618- 1746- 1874- 2002-  2129

Cell:   24 True Value:   1598.   Range:   1279.
         5.    1.    4.    3.    5.  111.    3.    1.    1.    4.
    851-  979- 1107- 1235- 1362- 1490- 1618- 1746- 1874- 2002-  2129
```

Synthetic Tabular Data – An Alternative to Complementary Cell Suppression

Ramesh A. Dandekar[1], Lawrence H. Cox[2]
[1]Energy Information Administration, U. S. Department of Energy,
Washington DC
[2]National Center for Health Statistics,
Centers for Disease Control and Prevention,
Hyattsville, MD
{Ramesh.Dandekar@EIA.DOE.GOV, LCOX@CDC.GOV}

Abstract: Complementary cell suppression is used for statistical disclosure limitation in tabular data, especially for magnitude data such as aggregate economic statistics. Cell suppression results in missing data, which complicates and can thwart thorough analysis. Suppressed entries can be replaced by interval estimates of their hidden values, but this too presents analytical challenges and can distort additivity to totals. Complementary cell suppression is an NP-hard computational problem. Even under optimal suppression, a data intruder can estimate expected values of suppressed entries, and often these estimates are close to original values. We introduce a new concept, *synthetic tabular data*, for limiting disclosure of sensitive information presented in tabular form. Synthetic tabular data is relatively easy to generate and provides significantly more information and flexibility than tables subject to suppression. The accuracy of synthetic cells is easy to control, making them useful for dissemination of statistical information.
Keywords: statistical disclosure limitation, statistical confidentiality

1. INTRODUCTION

Statistical disclosure occurs when released statistical data permit close approximation of sensitive information pertaining to an individual respondent or unit of analysis. A tabulation cell whose value closely approximates sensitive individual data is a *sensitive cell*. A cell is sensitive if its value equals the total for some statistic of data for only one or two respondents. Furthermore, if two respondents *dominate* the cell total, viz., the total contribution of all but the two largest contributors represents only a small fraction of the largest contribution, then the second largest can subtract its contribution from the cell total to obtain a narrow estimate of the largest. Values of sensitive cells must be *protected*, viz., obscured to the point that estimates of this sort of sensitive individual data are sufficiently imprecise. Federal Committee on Statistical Methodology (1994) provides an overview of statistical disclosure and disclosure limitation methods.

Procedures to protect sensitive cells in tabular data have evolved over the last four decades. From the very beginning national statistical offices realized that simply withholding the value for sensitive cells was insufficient to protect sensitive information in tables containing marginal totals. *Complementary cell suppression* (Cox 1980, 1995) was introduced and practiced by statistical offices to protect sensitive cells from disclosure through manipulation of additive relationships in statistical tables. Complementary cell suppression is aimed at assuring that *exact interval estimates* (lower and upper bounds) of the value of each suppressed sensitive cell are at a safe distance from the actual cell value, viz., lie within an interval at least as broad as that defined by predetermined *protection limits* (Cox 2001). (A generalization, *range protection*, allows protection limits to vary while enforcing a minimum distance between them.) In the largest-second largest contributor scenario, these limits equal a few percent of the largest contribution below and above the cell value. Cox (1981) provides a theory and algorithms for computing protection limits.

Early approaches to complementary suppression were based on linear equations (Fellegi 1972) and later linear programming (Sande 1984). Several approaches exploited properties of two-dimensional tables, e.g., assuring at least two suppressions in each row or column containing suppressions (Cox 1980) and network models for complementary cell suppression (Cox 1987, 1995), but, although efficient, such approaches do not generalize from two-dimensional to multi-dimensional tables or from simple hierarchies to complex aggregation structures (Cox and George 1989). Complementary cell suppression has been shown to be an *NP-hard problem* (Kelly et al. 1992), even for one-dimensional tables, making the existence of a computationally efficient, optimal method unlikely. Recent approaches are based on integer linear programs and branch-and-cut methods from integer programming (Fischetti and Salazar 2000).

Tables with suppressions are difficult to analyze. In lieu of suppressing cell values, it has been suggested, e.g., by Gordon Sande, that suppressions be replaced by their exact *interval estimates*. This is a step in the right direction, but is still demanding computationally and does not go sufficiently far in assuring ease-of-use for disclosure-limited tabular data. By using commonly practiced missing data techniques, e.g., *iterative proportional fitting* (Bishop et al. 1975) and the *E-M algorithm* (Little and Rubin 1987), probabilistic estimates for suppressed tabular cells can be computed, sometimes with great accuracy, sharply reducing the effectiveness of complementary cell suppression for statistical disclosure limitation. A third approach, introducing multiplicative noise into the underlying microdata (Zayatz et al. 2000), has been offered but not pursued.

In this paper, we demonstrate a new and different approach to limiting disclosure from sensitive tabular data cells. The method applies equally to two-dimensional tables as to multi-dimensional or linked tables, and to hierarchical as well as to complex tabular structure. We completely discard notions of complementary suppression and interval data and in their place advocate the use of *synthetic tabular data* to disseminate statistical information presented in tabular form. The essence of this approach is to replace each sensitive value with a value at a sufficient distance from the true value, and to adjust nonsensitive cell values minimally to restore additivity to totals. This method completely

eliminates information loss associated with complementary suppression procedures, restores analytical tractability, requires a fraction of the computational resources required by complementary cell suppression methods, and offers multiple alternative outputs through choice among several objective functions satisfying a wide variety of requirements meaningful to national statistical offices. This concept permits extension in various directions—theoretical, computational, and practical implementation. Examination of these opportunities is begun here.

Section 2 presents the underlying concept of synthetic tabular data and a precise mathematical formulation for the associated computational problem. This is a mixed integer linear programming problem involving binary variables. Because such problems are difficult to impossible to solve computationally, a heuristic is provided for assigning the binary variables, leaving only a linear program to be solved. Section 3 illustrates the method in three dimensions, and two examples based on complex linked tabular structure are presented in Section 4. Each of these examples is compared to an optimal or near-optimal suppression solution. The use and selection of heuristics is examined through extensive simulation in Section 5. The question of what it means to protect sensitive cells is reexamined in Section 6, leading potentially to less distortion of original data. Concluding comments are provided in Section 7.

2. SYNTHETIC TABULAR DATA: CONCEPT AND MATHEMATICAL FORMULATION

The objective in generating synthetic tabular data is to closely mimic the original tabular data, subject to obscuring sensitive cell values to a sufficient extent. The underlying concept is simple: The value of each sensitive cell is replaced by a synthetic value selected to be at a *safe distance* away from the true cell value. As a starting point, we set this distance to be either the sensitive cell's lower or its upper protection limit; alternatives are examined in Section 6. Some or all of the nonsensitive cell values are then adjusted from their true values by as small an amount as possible to restore additivity to totals within the tabular system.

Within our framework, adjustments to nonsensitive cell values can be controlled in various ways. Selected nonsensitive cells, e.g., zero cells, can be exempted from change. Adjustments can be confined to within meaningful limits such as sampling variability. One of several linear objective functions can be used to measure and assure minimum deviation.

Tabular data systems with marginal entries can be represented by their system of equations in matrix form: MX = 0. Column vector X represents the tabulation cells of the system; x* represents the original data. Matrix M is the *aggregation matrix* representing the tabular structure among the cells. The entries of M are -1, 0 or +1: each row of the M corresponds to one *aggregation* (tabular equation) in which "+1" denotes a contributing internal cell and "-1" a marginal cell. With this notation,

the mathematical structure of optimal synthetic tabular data is specified below by a mixed integer linear programming (MILP) formulation, analogous to that introduced in Cox (2000).

Notation

i = 1, …, p: denote the p sensitive cells
i = p+1,…, n: denotes the n-p nonsensitive cells
I_i = binary (zero/one) variable denoting selection of the lower/upper limit for sensitive cell i = 1,…,p
$LPROTECT_i$ = lower deviation required to protect sensitive cell i = 1,…,p
$UPROTECT_i$ = upper deviation required to protect sensitive cell i = 1,…,p
y_i^+ = positive adjustment to cell value i
y_i^- = negative adjustment to cell value i
UB_i, LB_i = upper/lower cell bounds on change to cell i
c_i = cost per unit change in cell i

MILP for Optimal Construction of Synthetic Tabular Data

$$\text{Min} \sum c_i \ (y_i^+ + y_i^-)$$

Subject to:

For i = 1,…, n:

$$M (y^+ - y^-) = 0$$
$$0 \leq y_i^+ \leq UB_i$$
$$0 \leq y_i^- \leq LB_i$$

For i = 1,…, p:

$$y_i^+ \geq LPROTECT_i * I_i$$
$$y_i^- \geq UPROTECT_i * (1 - I_i)$$

After solution of the MILP, the synthetic tabular data t = (t_i) is given by: $t_i = x*_i + y_i^+ - y_i^-$. Except as noted below, costs c_i are nonnegative, which implies that $y_i^+ y_i^- = 0$, viz., adjustment in a specific direction is indicated.

Five different cost functions are commonly used. They are: (1) constant, (2)log(1 + value), (3) value, (4) 1/(1+value), and (5) log(1+value)/(1+value), where 'value' denotes the cell value.
In general, mixed integer linear programming formulations are suitable only to solve small problems. We introduce a simple heuristic for selecting the binary I-variables, thereby reducing the problem to a linear programming formulation, which in practice can be efficiently solved for large and complex tabular structures.

The heuristic choice of assignment of sensitive cells to their lower/upper bound can be made in several ways. To illustrate our method, we introduce the following simple heuristic.

- Arrange all the sensitive cells in the table in an increasing order of magnitude of the cell values.
- Using an alternating sequence, assign value zero or one to the binary constant associated with each sensitive cell.
- When the marginal cell is sensitive and there are multiple internal sensitive cells, the direction of change of the marginal cell is reset to the net direction of change among the internal sensitive cells (when such exist).
- Any heuristic choice runs the risk of creating an infeasible problem. To ensure feasibility, we assign very high cost to adjustment of the sensitive cell in the opposite direction.

Other possible variations on the heuristic include: assign all sensitive cells to their lower (or upper bound), and, assign directions of change randomly. More complicated heuristics are also possible. In Section 5 we conduct a sensitivity analysis on the outputs based on these variations. As illustrated in Section 5, choice of heuristic appears to have minimal effect on quality and usefulness of the results.

3. ILLUSTRATION: THREE-DIMENSIONAL TABLE

We illustrate the method for a hypothetical three-dimensional table, containing 10 columns, 6 rows and 4 levels. Our table contains 191 non-zero cells, of which 24 cells are sensitive cells. It is customary, but not in all cases necessary, to exempt zero cells from change as, e.g., some zero cells are structural zeroes. We do so here. For simplicity, we assume *symmetric protection*, viz., $\mathrm{LPROTECT}_i = \mathrm{UPROTECT}_i = \mathrm{PROT}_i$. This is also customary.

The location of the sensitive cells, their cell values and required cell protection limits are illustrated in Table 1.

Table 1: Sensitive Cells and Protection Limits

Col	Row	Lev	Val	Prot	Col	Row	Lev	Val	Prot	Col	Row	Lev	Val	Prot
2	1	1	714	39	2	1	2	539	59	2	4	3	644	35
4	1	2	70	7	4	1	3	614	34	4	2	2	786	87
4	2	3	928	51	4	4	2	382	42	4	6	2	1238	17
5	1	1	140	7	6	2	2	1074	59	6	3	2	544	30
7	1	3	549	61	7	3	2	631	70	7	5	2	726	40
7	5	3	134	7	8	1	3	92	10	8	4	2	1050	58
8	5	1	664	36	8	5	4	664	36	9	2	1	1042	57
9	3	3	820	91	9	5	2	1598	88	9	5	4	1598	88

Using traditional complementary cell suppression techniques, following Kelly et al. (1992) and Zayatz (1992), our test example requires 39 complementary suppressions to protect 24 sensitive cells, displayed in Table 2. The complementary cells are marked by a symbol *c* next to the cell value, and the sensitive cells are marked by symbol ***w***. In addition, gray shades identify suppressed cells to emphasize the numeric values hidden from display. The complementary cell suppression in this example results in significant information loss, reducing the usefulness and usability of the table useless for many practical applications.

To generate a synthetic table that mimics Table 2 while limiting disclosure as specified in Table 1, we use the procedure described in Section 2. We choose costs equal to the cell value (3), which has the effect of targeting smaller nonsensitive cells for adjustment. This choice is arbitrary but in keeping with, e.g., past practice for U.S. Economic Censuses (Cox 1980, 1995). The cell value adjustments are such that resulting table is additive in all the dimensions and at the same time the published estimates for the sensitive cells are at one of the outer limits of their protection range.

Table 2: Cell Suppression—(10x6x4)Table

6764	714w	3356	4067c	140w	--	3932	1478c	--	20451
1994c	--	5593	--	3022	3504c	--	3220	1042w	18375
3744c	--	3708	--	3678c	2502c	--	--	--	13632
2810c	10632c	--	2445c	--	--	2313	2978	7548c	28726
3682	--	--	--	4667	1988c	1748	664w	--	12749
18994	11346	12657	6512	11507	7994	7993	8340	8590	93933
--	539w	--	70w	--	7472	715c	3832	--	12628
2253c	--	4948	786w	472	1074w	1830	5030	--	16393
640c	--	986	--	--	544w	631w	48c	750c	3599
1334c	--	1016	382w	3175	3302c	3803	1050w	--	14062
1648	2814	--	--	--	2102c	726w	--	1598w	8888
5875	3353c	6950	1238w	3647	14494	7705	9960	2348	55570
--	3552c	3476	614w	1916c	1131	549w	92w	1772	13102
--	--	3222	928w	--	--	308c	429	87c	4974
4145	--	--	3692	2115c	4196	414c	3804c	820w	19186
5995	644w	--	--	2410	1677c	--	1912c	4134c	16772
2016	--	--	2212	2826	1627c	134w	--	--	8815
12156	4196c	6698	7446c	9267	8631	1405	6237	6813	62849
6764	4805	6832	4751	2056	8603	5196	5402	1772	46181
4247	--	13763	1714	3494	4578	2138c	8679	1129c	39742
8529	--	4694	3692	5793	7242	1045c	3852c	1570c	36417
10139	11276	1016	2827	5585	4979	6116	5940	11682	59560
7346	2814	--	2212	7493	5717	2608	664w	1598w	30452
37025	18895	26305	15196	24421	31119	17103	24537	17751	212352

Table 3 summarizes the cell locations and magnitude of the *controlled adjustments* to true cell values. We have highlighted sensitive cells, in addition to marking them with symbol *w*, so that readers can easily verify that adjustments to sensitive cells are at either of their respective protection limits.

Table 3: Controlled Adjustments to (10x6x4)Table

--	39w	--	-41	-8w	--	--	10	--	--
--	--	--	--	18	13	--	26	-57w	--
--	--	--	--	8	-8	--	--	--	--
--	-35	--	-42	--	--	20	--	57	--
--	--	--	--	--	-5	41	-36w	--	--
--	4	--	-83	18	--	61	--	--	--
--	-68w	--	7w	--	--	61	--	--	--
--	--	--	87w	-18	-59w	--	-10	--	--
--	--	--	--	--	30w	-70w	-48	88	--
--	--	--	42w	--	-80	-20	58w	--	--
--	19	--	--	--	109	-40w	--	-88w	--
--	-49	--	136w	-18	--	-69	--	--	--
--	10	--	34w	8	--	-61w	-10w	19	--
--	--	--	51w	--	--	-164	-16	129	--
--	--	--	-33	-8	-22	70	84	-91w	--
--	35w	--	--	--	80	--	-58	-57	--
--	--	--	-105	--	-58	163w	--	--	--
--	45	--	-53	--	--	8	--	--	--
--	-19	--	--	--	--	--	--	19	--
--	--	--	138	--	-46	-164	--	72	--
--	--	--	-33	--	--	--	36	-3	--
--	--	--	--	--	--	--	--	--	--
--	19	--	-105	--	46	164	-36w	-88w	--
--	--	--	--	--	--	--	--	--	--

After applying the linear programming controlled adjustments to the original table, synthetic Table 4 results. Once again, we highlight the sensitive cells for ease of understanding. In a real application only the synthetic values are published. Depending on the accuracy of the data, statistical offices might attach to the cost function quality indicators designed to select cells of lower quality for adjustment, or for larger adjustment. Alternatively, the LB and UB could be based on sampling or measurement error. This is discussed further in Section 5.

In synthetic Table 4, true values are published for 106 cells. For the remaining 85 cells, published cell values are adjusted sufficiently from their true values to protect the sensitive cell values from disclosure within their protection interval. Most of the cell values of the marginal

cells are unaffected in the synthetic table, and the table is additive in all dimensions.

Table 4: Synthetic (10x6x4)Table

6764	753	3356	4026	132	--	3932	1488	--	20451
1994	--	5593	--	3040	3517	--	3246	985	18375
3744	--	3708	--	3686	2494	--	--	--	13632
2810	10597	--	2403	--	--	2333	2978	7605	28726
3682	--	--	--	4667	1983	1789	628	--	12749
18994	11350	12657	6429	11525	7994	8054	8340	8590	93933
--	471	--	77	--	7472	776	3832	--	12628
2253	--	4948	873	454	1015	1830	5020	--	16393
640	--	986	--	--	574	561	0	838	3599
1334	--	1016	424	3175	3222	3783	1108	--	14062
1648	2833	--	--	--	2211	686	--	1510	8888
5875	3304	6950	1374	3629	14494	7636	9960	2348	55570
--	3562	3476	648	1924	1131	488	82	1791	13102
--	--	3222	979	--	--	144	413	216	4974
4145	--	--	3659	2107	4174	484	3888	729	19186
5995	679	--	--	2410	1757	--	1854	4077	16772
2016	--	--	2107	2826	1569	297	--	--	8815
12156	4241	6698	7393	9267	8631	1413	6237	6813	62849
6764	4786	6832	4751	2056	8603	5196	5402	1791	46181
4247	--	13763	1852	3494	4532	1974	8679	1201	39742
8529	--	4694	3659	5793	7242	1045	3888	1567	36417
10139	11276	1016	2827	5585	4979	6116	5940	11682	59560
7346	2833	--	2107	7493	5763	2772	628	1510	30452
37025	18895	26305	15196	24421	31119	17103	24537	17751	212352

4. ILLUSTRATION: MULTI-DIMENSIONAL LINKED TABLES

The procedure of Section 2 for generating synthetic tabular data is applicable to all multi-dimensional or multi-dimensional linked tables. We next provide the overall performance statistics for synthetic tables based on two test examples of multi-dimensional linked tables.

The first test example consists of two five-dimensional linked sections of a six-dimensional table (6x4x16x4x4x4). The table contains 1254 non-zero cells. Of these, 1089 cells are nonsensitive and 165 cells are sensitive. Fischetti and Salazar (2000) determined that the optimum complementary cell suppression results in 419 suppressed cells, amounting to 34% of total non-zero cells.

The second example consists of four five-dimensional linked sections of a nine-dimensional table (4*29*3*4*5*6*5*4*5). The table contains 1141 non-zero cells, of which 831 cells are nonsensitive and 310 cells are sensitive. Fischetti and Salazar (2000) determined that the optimum complementary cell suppression results in 491 suppressed cells, which is 43% of total non-zero cells.

The synthetic tables generated by using these two test examples provide additive tables containing cell values for all the non-zero cells in the original test examples. In Table 5 we summarize the overall performance statistics of change from nonzero true value by ten different percent change from true value categories. We use five different cost functions that are commonly used in tabular cell protection to demonstrate five different possible formulations for synthetic tables.

From Table 5 it is clear that, by proper selection of the cost function, controlled adjustments could be targeted to specific nonsensitive cell categories. Irrespective of the choice of the cost function, approximately 75% of the nonzero cell values in the first test case and 50% of the nonzero cell values in the second test case are altered within less than 1% of their true cell value. The synthetic cells undergoing changes in excess of 5% of true cell value are typically sensitive cells, which are otherwise blocked from publication using the complementary cell suppression method.

The quality of cell-level information from the synthetic table could be conveyed to data users by using different strategies. As an option, a quality indicator, such as g (good), f (fair), and p (poor) could be assigned to each synthetic cell to inform the data user of the level of accuracy of information contained in each synthetic cell. Other options include: (1) providing overall percent accuracy of the published information, or (2) dividing the cells in multiple size categories and providing overall percent accuracy for each size category separately.

We have used only five basic cost functions to demonstrate the synthetic data generation technique in the linear programming environment. There is of course a wide spectrum of cost functions available to potential practitioner of synthetic tables. An advantage of the synthetic tabular framework is that with modest effort several approaches could be tried and the “best” selected.

Table 5: Number of Cells by Percent Change[1]

2 Sections Of Six-Dimensional Linked Table

Percent change from true value	Cost Function Used In Optimization				
	constant	log(value)	value	1/value	log(value)/value
.00- .10	691{ 55.3%}	716{ 57.5%}	749{ 60.4%}	720{ 57.5%}	687{ 54.8%}
.10- .50	189{ 70.4%}	154{ 69.8%}	120{ 70.1%}	231{ 75.9%}	254{ 75.1%}
.50- 1.00	91{ 77.7%}	72{ 75.6%}	37{ 73.1%}	47{ 79.6%}	56{ 79.6%}
1.00- 1.50	38{ 80.7%}	27{ 77.8%}	41{ 76.4%}	22{ 81.4%}	28{ 81.8%}
1.50- 2.00	22{ 82.5%}	33{ 80.4%}	22{ 78.1%}	14{ 82.5%}	14{ 82.9%}
2.00- 5.00	52{ 86.6%}	52{ 84.6%}	63{ 83.2%}	47{ 86.3%}	42{ 86.3%}
5.00- 10.00	73{ 92.5%}	88{ 91.7%}	98{ 91.1%}	119{ 95.8%}	100{ 94.3%}
10.00- 15.00	58{ 97.1%}	56{ 96.1%}	51{ 95.2%}	51{ 99.8%}	69{ 99.8%}
15.00- 30.00	19{ 98.6%}	24{ 98.1%}	30{ 97.7%}	2{100.0%}	3{100.0%}
30.00-100.00	17{100.0%}	24{100.0%}	29{100.0%}	0{100.0%}	0{100.0%}
Unchanged cells	390{ 31.2%}	422{ 33.9%}	651{ 52.5%}	319{ 25.5%}	257{ 20.5%}

4 Sections Of Nine-Dimensional Linked Table

Percent change from true value	cost function used for optimization				
	const	log(value)	value	1/value	log(value)/value
.00- .10	431{ 38.1%}	397{ 35.1%}	494{ 44.0%}	320{ 29.3%}	333{ 29.9%}
.10- .50	96{ 46.6%}	134{ 46.9%}	33{ 46.9%}	46{ 33.5%}	69{ 36.1%}
.50- 1.00	59{ 51.8%}	48{ 51.2%}	27{ 49.3%}	23{ 35.6%}	46{ 40.3%}
1.00- 1.50	35{ 54.9%}	23{ 53.2%}	29{ 51.9%}	23{ 37.7%}	27{ 42.7%}
1.50- 2.00	33{ 57.8%}	29{ 55.8%}	13{ 53.0%}	25{ 40.0%}	15{ 44.0%}
2.00- 5.00	85{ 65.3%}	90{ 63.7%}	86{ 60.7%}	83{ 47.6%}	90{ 52.1%}
5.00- 10.00	256{ 87.9%}	259{ 86.6%}	212{ 79.5%}	242{ 69.7%}	266{ 76.0%}
10.00- 15.00	55{ 92.8%}	64{ 92.3%}	57{ 84.6%}	60{ 75.2%}	62{ 81.6%}
15.00- 30.00	32{ 95.6%}	45{ 96.3%}	58{ 89.8%}	81{ 82.6%}	59{ 86.9%}
30.00-100.00	50{100.0%}	42{100.0%}	115{100.0%}	190{100.0%}	146{100.0%}
unchanged cells	353{ 31.2%}	329{ 29.1%}	453{ 40.3%}	287{ 26.3%}	302{ 27.1%}

[1] The numbers in the parentheses are cumulative percentages associated with the cell count.

5. USE AND SELECTION OF A HEURISTIC

A precise mathematical formulation for generating synthetic tabular data, as a mixed integer linear program, was provided in Section 2. Also in Section 2, we proposed replacing optimal selection of direction for change of sensitive cells (the integer portion of the MILP) by a simple heuristic, thus reducing the computational problem to a linear program. It is appropriate to examine two questions:

- **Is optimal selection of direction for change of sensitive cells necessary, or, can a heuristic be used?**
- **How does this heuristic compare with other potential heuristics?**

5.1 Optimal Vs. Heuristic Selection of Direction for Change

If a mathematical optimization is computable, the optimization will produce one or more solutions that are provably "best" with respect to the constraints and objective function(s) specified in the mathematical formulation. The purpose of constructing an optimal solution is not, however, necessarily its actual use. Mathematical constraints typically only approximate real-world conditions. Mathematical formulations typically incorporate only a subset of actual conditions and criteria, and often are only approximations, with the result that optimal solutions only approximate fully "best" solutions. Likewise, two solutions that differ in objective function value for practical purposes are often indistinguishable. In many situations, therefore, demonstration of an optimal solution is valuable primarily from the standpoint of establishing a "gold standard" against which other solutions or outcomes can be compared.

This is true in the synthetic data framework. An optimal solution to the MILP of Section 2 does not necessarily exhibit distributional properties identical to those of the original data, and therefore is not guaranteed to produce equivalent results for every conceivable statistical analysis. (This, incidentally, is equally, if not more, true for cell suppression or interval data.) Conversely, a synthetic data set that, say, is within measurement error of original data is arguable equivalent to the original, regardless of objective function value. The mathematical constraints and objective function specified in Section 2 are designed to produce a synthetic result close to original data, but at some point there is no practical distinction between two similar solutions. Consequently, a fully optimal solution is not required to generate useable synthetic tabular data.

How then to proceed? Based on sampling and other measurement error, an estimated standard error can be computed for each tabulation cell. Within our linear programming model, it is a simple matter to further constraint the controlled adjustments (y-variables) to within, say, two standard errors of original data. Any two such solutions differing by no more than two standard errors are for all practical purposes equivalent. Using an

appropriate heuristic to select direction of change, run the linear program. If at least one feasible solution exists, then an acceptable synthetic tabular data set has been found. In general, the relatively large number of nonsensitive cells will ensure feasibility. In the next subsection, we examine and compare different possible choices of heuristic.

5.2 Effect of Choice of Heuristic

A simple heuristic for selecting directions of change for sensitive cells was presented in Section 2, based on sorting the sensitive cells and assigning lower/upper protection to each in an alternating manner. Other heuristics are possible. In this subsection we illustrate and compare selection heuristics.

There are several obvious choices, including: the alternating heuristic of Section 2, referred to as "Plus/Minus"; for each sensitive cell, selecting the lower bound direction (viz., I = 0), referred to as "Minus"; for each sensitive cell, selecting the upper bound direction (viz., I = 1), referred to as "Plus"; and, for each sensitive cell, selecting the direction randomly, simulated 100 times. The evaluation statistics are: total *change* (controlled adjustments); total of original cell values affected by change; average change by value; number of cell values changed; average percent change in cell value; and, total percent change in cell value. The results, based on Table 3, are presented in Table 6.

Table 6 : Comparison of Heuristics for Table 3
("Change" measured by absolute value)

Comparison of Plus/Minus, Minus and Plus Heuristics

	Quantity Changed (1)	Affected Quantity (2)	Average Change	Number of Cells	Average % Change	Tot.%Chng. (=(1)÷(2))
Plus/Minus	4364.	221980.	51.34118	85	8.63305	1.96594
Minus	4460.	177172.	58.68421	76	8.76424	2.51733
Plus	4370.	210424.	52.65060	83	7.61722	2.07676

Random Selection of Direction—Statistics for 100 Simulations

	Quantity Changed (1)	Affected Quantity (2)	Average Change	Number of Cells	Average % Change	Tot.%Chng. (=(1)÷(2))
Mean	4046.	217252.	47.92028	85	6.95427	1.87373
Std. Dev.	431.	18767.	4.94445	4	1.35592	.23795
Min.	3058.	168143.	38.02299	73	3.76409	1.36656
Max.	5336.	264115.	62.77647	93	10.68496	2.55154

The first half of Table 6 reveals that the base heuristic works slightly better than the two extreme choices. The second half of the table provides statistics on 100 simulations in which the magnitude of protection level for sensitive cells was exactly the same as the base case, but the direction of adjustment to sensitive cells was selected randomly. Based the mean and standard deviation over the 100 trials, it does not appear that random selection offers measurable improvement over the base case. Moreover, minimum values associated with all six "statistical change" measures were associated with six different simulations. Furthermore, none of the 100 offered convincing improvement.

From these modest analyses, we conclude that it is unlikely that a "best" heuristic can be found. Indeed, this actually is a strength of the synthetic tabular framework, because the relatively low computational cost associated with producing one or more sets of synthetic tabulations with respect to a single heuristic facilitates experimentation with multiple heuristics. The "best" simulated data set can then be selected from an array of candidates based on appropriate criteria including expert judgment.

6. INTERPRETING CONFIDENTIALITY PROTECTION IN THE SYNTHETIC DATA CONTEXT

Synthetic tabular data alters original data. The degree of distortion is determined by the number of sensitive cells and required changes to sensitive cell values. Based on the cell suppression paradigm, in the model of Section 2 these changes are set equal to the protection deviations PROT_i, viz., each sensitive value is forced to one of its protection limits. This is necessary under cell suppression because allowing estimation of the cell value within a narrower range is by definition not permissible. However, a more flexible interpretation of protection is possible in the synthetic data framework, as follows.

If a tabulation cell represents data from only one respondent, then the cell value is a point estimate of the contribution of the respondent. It would be unwise to select a synthetic value too close to the true value, and therefore use of PROT_i is appropriate. Similarly, if the cell contains data for precisely two respondents, then either can subtract its value from the published cell value and use the result as a point estimate of the contribution of the other. Therefore, full protection makes sense in this situation as well.

However, when a small number of respondents (but more than two) dominate the cell value, the disclosure problem for synthetic tabular data is less clear, as illustrated by the following example.

Assume that disclosure is defined as allowing the second largest to estimate the contribution *c* of the largest to within k-percent of its value. Given a sensitive cell with largest contribution *c* and second largest contribution *d,* assume that the total contribution *e* of the remaining respondents (respondents 3, 4, … etc.) equals q-percent of the largest contribution, viz., $e = c(q/100)$ with $q < k$. Then, from Cox (1981), $\mathrm{PROT}_i = c(k - q)/100$. A synthetic value *s* is published in lieu of the true cell value $c + d + e$. The second largest contributor (the *intruder*) subtracts its contribution *d* from synthetic value *s*, obtaining a point estimate $s - d$ of the contribution of the largest. This estimate is imprecise, for two reasons. First, the intruder cannot account precisely for the total contribution *e* of the remaining respondents. Second, the intruder does not know whether the synthetic value *s* lies below or above the true cell value, or how close. Even assuming that the intruder can estimate *e* to within 100-perecent of its value, viz., within the interval [0, 2e], the intruder still only has a range of point estimates [s − d − 2e, s − d] of *c* that may not even contain the actual contribution *c*.

This makes it reasonable to consider relaxing the requirement to force each synthetic sensitive value all the way to one of its protection limits. This clearly is a policy decision, requiring further analysis based on actual sensitive data. To illustrate the effects of this relaxation, we simulated going only "half-way" in Table 3. Namely, having selected the direction of change for a sensitive cell value using the Minus/Plus heuristic, we randomly select the adjustment to sensitive cell *i* within the range [$\text{Prot}_i/2$, Prot_i] using a uniform distribution, simulated 100 times. The results are presented in Table 7.

Table 7: Smaller Protection Level Selected Randomly—100 Simulations (NewProt_i = Uniform [$\text{Prot}_i/2$, Prot_i]); Direction Random as in Table 3)

	Quantity Changed	Affected Quantity	Average Change	Number of Cells	Average % Change	Tot.%Chng. (=(1)÷(2))
Mean	3429.	214193.	40.71679	84	6.82568	1.60473
Std. Dev	236.	10677.	3.29047	3	.73915	.13342
Min.	2866.	186637.	33.71765	77	5.17424	1.33369
Max.	3952.	237620.	51.32468	91	8.61083	1.97567

Comparing Table 7 with the first row of Table 6, it is clear that less distortion results, with protection still assured.

7. CONCLUDING COMMENTS

Synthetic tabular data offers a more attractive option for disseminating tabular data containing sensitive information than conventional complementary cell suppression. Complementary cell suppression results in a significant amount of information loss, irrespective of how close one gets to optimum suppression. The overall information generated by complementary cell suppression fails to compare favorably to synthetic tabular data both in completeness and usability. Complementary cell suppression is a computationally demanding, and optimal suppression is an NP-hard problem, whereas the computational effort required to generate synthetic tables is minimal. This allows the statistical office to generate multiple synthetic data scenarios and select the most favorable based, among other criteria, on expert judgment.

In this paper we introduced the concept of synthetic tabular data and provided a simple heuristic combined with linear programming methods for generating synthetic tabular data. Illustrations for multi-dimensional and linked tables were provided. Alternatives for selecting direction for change were examined and compared. A more flexible interpretation of confidentiality protection in tabular data was examined.

Computational techniques, such as iterative proportional fitting and the EM algorithm, could also be used to generate synthetic tabular data. Such methods are useful, e.g., when all internal cells are suppressed or unavailable and must be estimated from marginal totals. However, in actual practice, not all marginal totals are fixed and such methods are likely to provide estimates unacceptably close to sensitive cell values.

Heuristics presented in this paper could be extended or replaced. In general, and for actual purposes, however, the methods presented here will result in practical, usable tabular data, and provide a basis for specialized approaches tailored to particular data. We compared several reasonable computational heuristics and found that they produced essentially equivalent results.

Having established a conceptual, practical and computational basis for synthetic tabular data, we examined the question of what constitutes adequate protection for a sensitive cell. In the synthetic data setting, a more flexible, data-enhancing interpretation emerged. This will require further practical simulation and examination from a policy standpoint by statistical offices.

In summary, synthetic tabular data reproduces original data as closely as possible, subject to confidentiality requirements, and offers considerable flexibility for preserving original values and for providing disclosure protection at less cost in terms of computational requirements and distortion of true values. The synthetic tabular framework offers advantages both to data producers and data users not possible under the more restrictive complementary cell suppression regimen.

DISCLAIMER

The material presented herein has been reviewed and approved by the Centers for Disease Control and Prevention for publication. It is solely the work 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

Bishop, Y, S. Fienberg and P. Holland (1975), Discrete Multivariate Analysis—Theory and Practice, Cambridge, MA: MIT Press.

Cox, L.H. (1980), "Suppression Methodology and Statistical Disclosure Control," *Journal of the American Statistical Association* 75, 377-385.

______ (1981), "Linear Sensitivity Measures in Statistical Disclosure Control," *Journal of Statistical Planning and Inference* 5, 153-164.

______ (1987), "New Results in Disclosure Avoidance for Tabulations," Bulletin of the International Statistical Institute, Proceedings of the 46th Session, Voorburg: International Statistical Institute, 83-84.

_____ (1995), "Network Models for Complementary Cell Suppression," *Journal of the American Statistical Association* 90, 1453-1462.

_____ (2000), "Discussion (of Session 49: Statistical Disclosure Control for Establishment Data)," ICES II: The Second International Conference on Establishment Surveys—Survey Methods for Businesses, Farms and Institutions, Alexandria, VA: American Statistical Association, 905-907.

_____ (2001), "Disclosure Risk for Tabular Economic Data," Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies (P. Doyle, J. Lane, J. Theeuwes and L. Zayatz, eds.), Chapter 8, New York: Elsevier, 2001, 167-183.

_____ and J. George (1989), "Controlled Rounding for Tables with Subtotals," *Annals of Operations Research* 20, 141-157.

Federal Committee on Statistical Methodology (1994), Statistical Policy Working Paper 22: Report on Statistical Disclosure and Statistical Disclosure Limitation Methodology, Washington, DC: U.S. Office of Management and Budget.

Fellegi, I. (1972), "On the Question of Statistical Confidentiality," *Journal of the American Statistical Asociation* 67, 7-18.

Fischetti, M. and J. J. Salazar (2000), "Models and Algorithms for Optimizing Cell Suppression Problem in Tabular Data with Linear Constraints", *Journal of the American Statistical Association* 95, 916-928.

Kelly, J., B. Golden and A. Assad (1992), "Cell Suppression: Disclosure Protection for Sensitive Tabular Data," *Networks* 22, 397-417.

Little, R. and D. Rubin (1987), Statistical Analysis with Missing Data, New York: John Wiley and Sons, Inc.

Sande, G. (1984), "Automated Cell Suppression to Preserve Confidentiality of Business Statistics," *Statistical Journal of the United Nations ECE* 2, 33-41.

Zayatz, L. (1992), "Using Linear Programming Methodology for Disclosure Avoidance Purposes", Bureau of the Census Research Report Series no. RR-92/02, Washington, DC: Bureau of the Census.

_____, T. Evans and J. Slanta (2000), "Using Noise for Disclosure Limitation of Tabular Establishment Data," ICES II: The Second International Conference on Establishment Surveys—Survey Methods for Businesses, Farms and Institutions, Alexandria, VA: American Statistical Association, 877-886.

February 27, 2002