WP. 41
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Geneva, Switzerland, 9-11 November 2005)

Topic (vi): Software for statistical disclosure control

COMPLEMENTARY CELL SUPPRESSION SOFTWARE TOOLS FOR STATISTICAL DISCLOSURE CONTROL – REALITY CHECK

Invited Paper

Submitted by the U.S. Department of Energy, United States of America[1]

[1] Prepared by Ramesh A. Dandekar (ramesh.dandekar@eia.doe.gov).

Complementary Cell Suppression Software Tools for Statistical Disclosure Control - Reality Check

Ramesh A. Dandekar
Statistics and Methods Group, U. S. Department of Energy, Washington DC
(Ramesh.Dandekar@EIA.DOE.gov -- http://mysite.verizon.net/vze7w8vk/)

1 Introduction

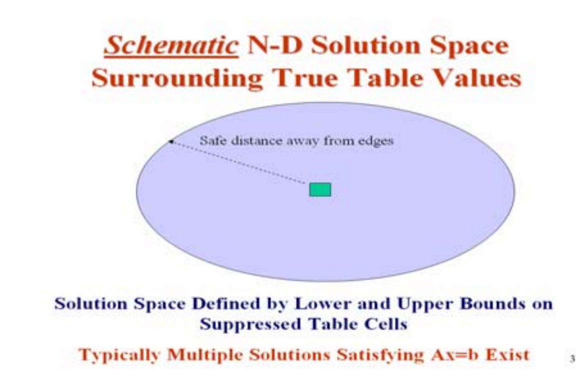
Currently, complementary cell suppression software tools are mostly used by statistical agencies to protect sensitive tabular data from disclosure. It is generally believed that the linear programming (LP) based complementary cell suppression procedures offer the best protection from wrongful disclosure of statistical information. In recent years LP-based cell suppression auditing software tools 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 auditing software tools. We use widely employed 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.

2 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.

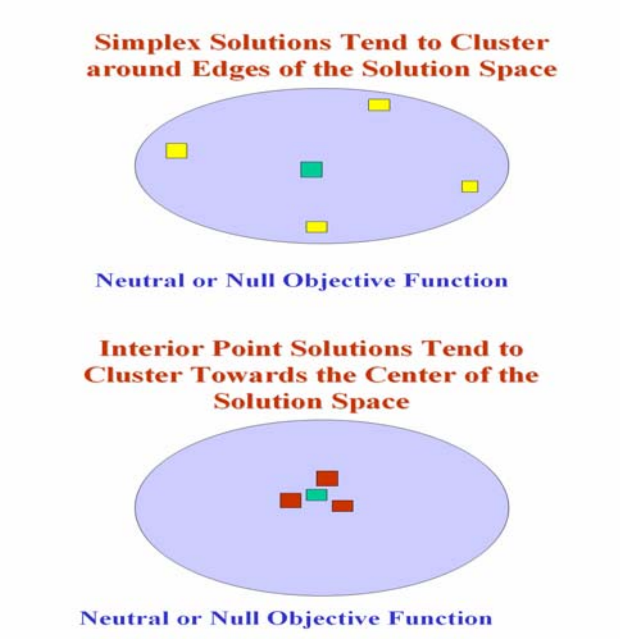
3 Current Tools

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 $xj = 0$.



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.

4 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".

5 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. Based on Table 4 statistics, 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.

6 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}.

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 to ensure additivity of the tabular cells

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

7 Conclusion 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

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 Bound	True Value	Upper Bound	percent Lower	percent 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:

	I	J	K	Desired Prot	Value 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
```