(In)Effectiveness of Independent Rounding of Discrete Tabular Data as Statistical Disclosure Control Strategy

Ramesh A Dandekar
EI-70, U. S. Department of Energy, Washington DC

Abstract: In an attempt to protect sensitive counts data (discrete data), independent rounding of tabular data cells has been proposed and has also been implemented by statistical agencies all over the world. In this paper we demonstrate that such a practice not only results in to (1) degradation of tabular data quality and (2) produces non-additive tables, but the strategy also fails to provide adequate protection from statistical disclosure to low count tabular data cells.

Introduction:

It has been a common practice to use independent rounding in public use (discrete and magnitude) tabular data. Independent rounding to the base 10 or its multiple has been commonly used, mostly for cosmetic reasons, by statistical agencies all over the world for a long time.

In recent years, however, rounding to base 3 or 5 has been proposed widely as a potential disclosure control strategy for a public use discrete tabular data products resulting from population census. The countries proposing to use this technique, based on the technical papers presented at UNECE2009 statistical disclosure control workshop (http://www.unece.org/stats/documents/2009.12.confidentiality.htm), include:

- United States of America
- United Kingdom
- Australia
- New Zealand
- European Union (EU) Countries

Contrary to current belief, in this paper we demonstrate that the independent rounding of discrete tabular data does have a significant potential for individual identification or inadvertent statistical disclosure of relatively low count table cells imbedded in tabular format data. We also show that readily available public domain linear programming (LP) software such as LP_SOLVE, BPMPD, HOPDM, PcX etc. as well as Disclosure Audit Software (DAS) developed by the Federal Committee of Statistical Methodology (FCSM) to ensure adequacy of complementary cell suppression pattern can also be used to identify potential disclosures resulting from the independent rounding process. The DAS software is available for use in a SAS environment.

Generic Independent Rounding Procedure:

Multiple variations of independent rounding procedures have been proposed in literature and have been practiced by statistical agencies all over the world. These procedures can be summarized by using the following generic notations:

Random Rounding to integer base b

x = cell value prior to rounding
k = largest integer multiplier of base b such that bk < x
r = residual value x – bk
x is rounded up to b(k+1)………. if r/b > 0.5
x is rounded down to bk………. if r/b < 0.5

Illustrative Example:

In this paper we use the procedure which is a special case of generic independent rounding procedure described above and is documented for actual use at the web site http://trbcensus.com/drb/03122008.pdf The procedure is used widely by the United States Census Bureau since 2001 to protect special tabulations resulting from the American Community Survey (ACS). The rounding procedure and the relevant text associated with the rounding instructions are summarized in the census document as shown below:

0 REMAINS 0
1-7 ROUNDS TO 4
8 OR GREATER ROUNDS TO NEAREST MULTIPLE OF 5

Any totals or subtotals needed should be constructed before rounding. This assures that universes remain the same from table to table, and it is recognized that cells in a table will no longer be additive after rounding.

The detailed background for the published tables resulting from this procedure is available from http://www.fhwa.dot.gov/ctpp/ . The tabular data is used by the U. S. Department of Transportation (DOT) and by state and local transportation agencies to generate census transportation planning products (CTPP). The further details on CTPP data products are available from http://www.fhwa.dot.gov/ctpp/dataprod.htm .

Problem Identification:

To demonstrate the effect of the independent rounding rule such as the one used by the census bureau for CTPP, we have used a simple hypothetical 5x4 table with small counts data as shown in table1. In this table, the last column is a sum of first four columns and the last row is sum of first three rows. The table grand total could be obtained by adding the last column or the last row associated with internal table cells. As a result, the table is additive in all respects.

Table 2 shows the outcome after census's independent rounding rules developed for CTPP are applied to the values of all cells in table1 in stand-alone mode. As can be seen from the table 2, after the independent rounding procedure is implemented, the table additivity is completely lost.

In table 3 we have summarized the overall quality aspects (lack of) of the rounded cell values from table 2 by comparing them with the true values presented in table 1. The (lack of) quality outcome from table 3 is similar to the one observed by the actual CTPP data users of independently rounded tables which is discussed in great details in a technical report available from the website **http://www.fhwa.dot.gov/ctpp/balance.htm**

Table 1

ILLUSTRATIVE EXAMPLE: ORIGINAL COUNTS TABLE

7	7	7	7	28
6	6	6	6	24
5	5	5	5	20
18	18	18	18	72

Ramesh A Dandekar, JSM2010,
Vancouver, Canada

5

Table 2

ILLUSTRATIVE EXAMPLE
PUBLISHED TABLE AFTER APPLYING DRB RULES

4	4	4	4	30
4	4	4	4	25
4	4	4	4	20
20	20	20	20	70

Original Values

7	7	7	7	28
6	6	6	6	24
5	5	5	5	20
18	18	18	18	72

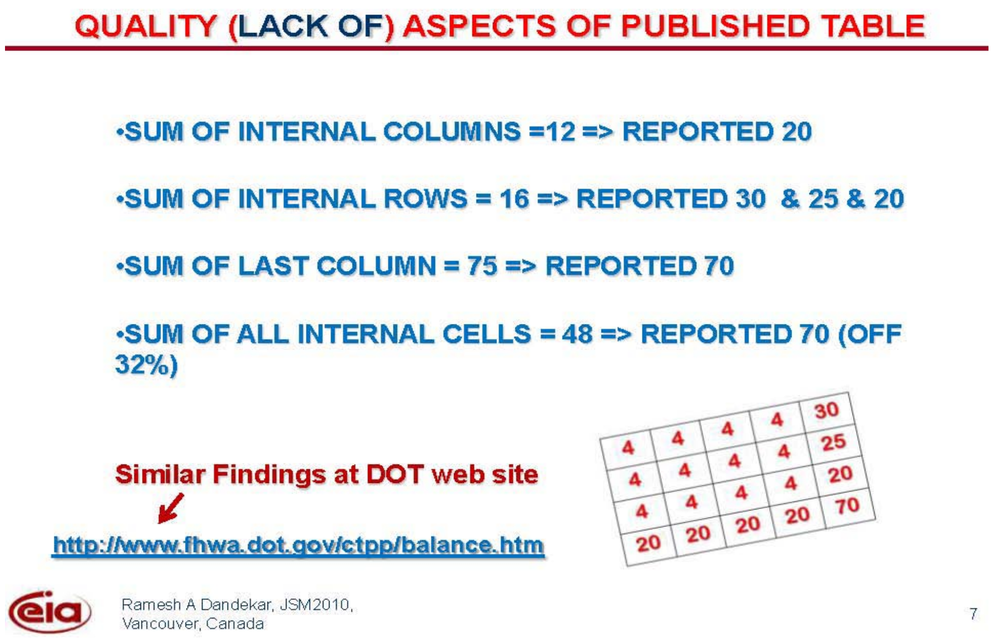
USCB Method:

0 REMAINS 0
1-7 ROUNDS TO 4
8 OR GREATER ROUNDS TO NEAREST MULTIPLE OF 5

Ramesh A Dandekar, JSM2010,
Vancouver, Canada

6

Table 3



Restoring Additive Properties of Independently Rounded Tables:

The non-additive nature of the independent rounding procedure is readily acknowledged by the staff at the U. S. Census Bureau in its correspondence with the CTPP data users. This leaves CTPP data users without any potential solution in case the table additivity is the primary concern.

We, however, observe that the routinely used linear programming procedures that are familiar to professionals with the operations research and mathematical optimization background could be implemented for use in CTPP environment to restore the additivity of the census published tables with relative ease.

In table 4, we show the re-created additive table created by using the linear programming technology. The LP setup uses publicly available rounding rules in combination with additive properties associated with the table structure. It is worth noting that as long as the rounding rules are strictly followed in creating public use tables, it is *always* possible to recreate additive table cell values. Implied table cell bounds, in combination with narrow ranges for these bounds (plus or minus 2 in our case) allows individual with linear programming expertise to determine the actual theoretical bounds (upper and lower bound estimates) for each published table cell value.

Table 4

RECREATED TABLE AFTER APPLYING LP-BASED TABULAR ADJUSTMENTS

7	7	7	7	28
7	5	7	7	26
4	6	4	4	18
18	18	18	18	72

IT IS ALWAYS POSSIBLE TO RECREATE ORIGINAL CELL VALUES WITH HIGH PRECISION USING LP TECHNOLOGY

Original Values

7	7	7	7	28
6	6	6	6	24
5	5	5	5	20
18	18	18	18	72

Ramesh A Dandekar, JSM2010,
Vancouver, Canada

9

In table 5, we show a typical linear programming data input file that could be used to solve the linear programming problem when table additivity is the prime concern for data user. The input file is compatible with LP_SOLVE linear programming solver available in the public domain. The solver is available from http://www.cs.sunysb.edu/~algorith/implement/lpsolve/implement.shtml and from many other web sites. As can be seen, the LP input requirements are minimal. Table additivity constraints and actual table cell bounds from table originator/creator/publishers is all that is required.

Table 5

ILLUSTRATIVE EXAMPLE LP_SOLVE FORMAT INPUT

```
•MAX: +1 W01;
•ROW01: +W01 +W04 +W07 +W10 -W13 = 0;
•ROW02: +W02 +W05 +W08 +W11 -W14 = 0;
•ROW03: +W03 +W06 +W09 +W12 -W15 = 0;
•ROW04: +W16 +W17 +W18 +W19 -W20 = 0;
•ROW05: +W01 +W02 +W03 -W16 = 0;
•ROW06: +W04 +W05 +W06 -W17 = 0;
•ROW07: +W07 +W08 +W09 -W18 = 0;
•ROW08: +W10 +W11 +W12 -W19 = 0;
•ROW09: +W13 +W14 +W15 -W20 = 0;
•W01 >   1 ;
•W01 <   7 ;
•W02 >   1 ;
•W02 <   7 ;
•W03 >   1 ;
•W03 <   7 ;
•W04 >   1 ;
•W04 <   7 ;
•W05 >   1 ;
•W05 <   7 ;
•W06 >   1 ;
•W06 <   7 ;
•W07 >   1 ;
•W07 <   7 ;
•W08 >   1 ;
•W08 <   7 ;
•W09 >   1 ;
•W09 <   7 ;
•W10 >   1 ;
•W10 <   7 ;
W11 >   1 ;
W11 <   7 ;
W12 >   1 ;
W12 <   7 ;
W13 >   28 ;
W13 <   32 ;
W14 >   23 ;
W14 <   27 ;
W15 >   18 ;
W15 <   22 ;
W16 >   18 ;
W16 <   22 ;
W17 >   18 ;
W17 <   22 ;
W18 >   18 ;
W18 <   22 ;
W19 >   18 ;
W19 <   22 ;
W20 >   68 ;
W20 <   72 ;
```

W01 1-7	W04 1-7	W07 1-7	W10 1-7	W13 28-32
W02 1-7	W05 1-7	W08 1-7	W11 1-7	W14 23-27
W03 1-7	W06 1-7	W09 1-7	W12 1-7	W15 18-22
W16 18-22	W17 18-22	W18 18-22	W19 18-22	W20 68-72

Ramesh A Dandekar, JSM2010,
Vancouver, Canada

11

The first line in the LP input file from table 5 is used to specify an objective function which could either be a minimization problem or a maximization problem. The variable specified in the objective function is used as criteria to satisfy the equality constraints and the bounds used to create original table.

In table 6, we show the output from the LP_SOLVE program which provides additive table cell values when maximizing criterion for variable W01 (cell in row 1 and column 1) in the table is used. As can be seen from this example, the estimated values more often than not are same as the original table cell values prior to rounding.

Table 6

LP_SOLVE OUTPUT

- Value of objective function: 7
- W01 7
- W02 7
- W03 4
- W04 7
- W05 5
- W06 6
- W07 7
- W08 7
- W09 4
- W10 7
- W11 7
- W12 4
- W13 28
- W14 26
- W15 18
- W16 18
- W17 18
- W18 18
- W19 18
- W20 72

ESTIMATED

7	7	7	7	28
7	5	7	7	26
4	6	4	4	18
18	18	18	18	72

eia Ramesh A Dandekar, JSM2010, Vancouver, Canada 12

In linear programming format our example table is represented by 20 unknown variables (Wxx, xx= 1, 20). By using input structure from table 5 and by alternately changing optimization criteria to minimize and maximize for all the 20 table unknown Wxx, one can potentially generate 40 different additive tables. By using the outcome from these 40 tables, one can determine the distribution of cell value, as well as the theoretical bounds for each table variable. In case where the theoretical upper and lower bound for a table is identical, the exact true table cell value can be estimated. If the cell for which the exact value can be determined is sensitive cell, it creates a potential individual privacy violation.

The linear programming input above uses the same logic and similar format implemented in statistical disclosure audit software routinely used by many statistical agencies. In the recent years DAS developed by the FCSM is available from the web site **http://www.fcsm.gov/committees/cdac/DAS.html** and is advocated for such a use.

As a potential disclosure control strategy, statistical agencies that continue to use independent rounding of discrete tabular data should use the DAS software to evaluate rounded table cell values for inadvertent disclosure of sensitive cell values.

With Appropriate modifications to the DAS software, statistical agencies may want to look at the feasibility of providing bounded and additive table cell values which offer adequate distortion of sensitive cells.

Summary Conclusions:

Based on the simple demonstration in this paper, the following conclusions could be derived.

- Current statistical disclosure control strategy, which relies on independent rounding of discrete tabular data is not safe and unnecessarily degrades data quality.
- It is always possible to recreate additive table cell values by using the basic information available in rounded discrete value tables.
- Recreated additive tables create a potential for a statistical as well as exact disclosure of sensitive table cell values.
- A safer disclosure control strategy would be to create additive, bounded and well protected discrete table cell values by using a linear programming method like the one outlined by Dandekar (2001). An additional benefit of this methodological change would be higher quality data dissemination.

References

Dandekar R. A. (2001) "Synthetic Tabular Data: A Better Alternative To Complementary Data Suppression - Original Manuscript Dated December 2001". Energy Information Administration, U. S. Department of Energy._ Also available from CENEX-SDC Project International Conference, PSD2006, Rome, Italy, December 13-15, 2006, Companion CD Proceedings ISBN: 84-690-2100-1.

Dandekar R. A. and Cox L. H. (2002), Synthetic Tabular Data: An Alternative to Complementary Cell Suppression, 2002. Manuscript, Energy Information Administration, U. S. Department of Energy.

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)

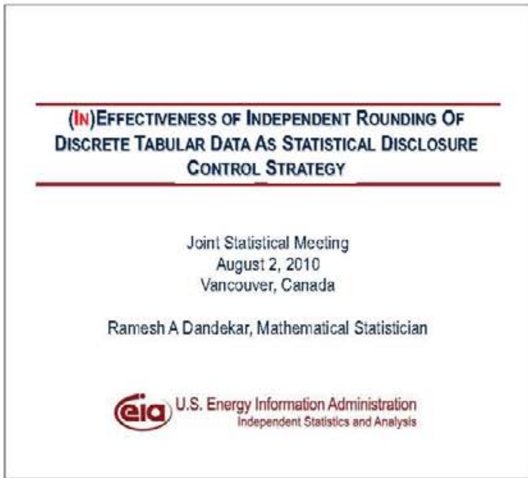
Dandekar Ramesh A. (2004), Maximum Utility-Minimum Information Loss Table Server Design for Statistical Disclosure Control of Tabular Data, pp 121-135, Lecture Notes in Computer Science, Publisher: Springer-Verlag Heidelberg, ISSN: 0302-9743, Volume 3050 / 2004**,** Title: Privacy in Statistical Databases: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

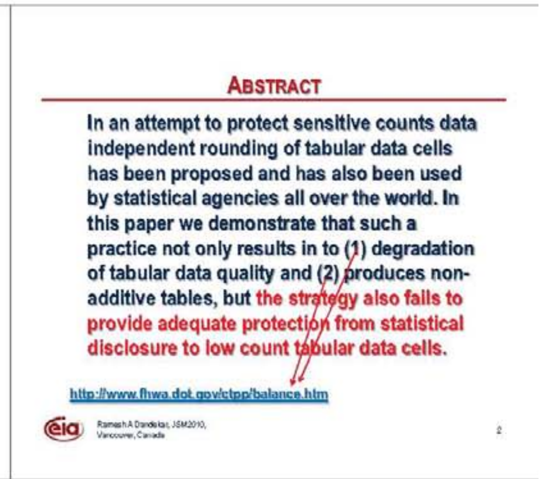
Dandekar Ramesh A. (2007), "Comparative Evaluation of Four Different Sensitive Tabular Data Protection Methods Using a Real Life Table Structure of Complex Hierarchies and Links", working paper 17, UNECE Work session on statistical data confidentiality (Manchester, United Kingdom, Dec 17-19, 2007)

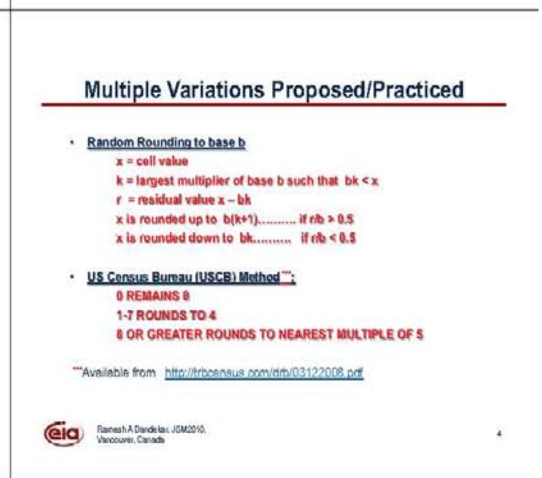
Dandekar Ramesh A. (2009), "Statistical Disclosure Control Of Tabular Format Magnitude Data - Why It Is Not A Good Idea To Use Home Grown Cell Suppression Procedures", Presented At FCSM2009 Conference.

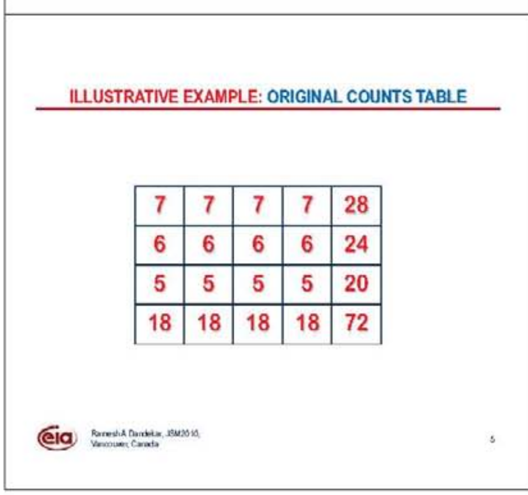
Dandekar Ramesh A. (2009), "Incorporating Quality Measures in Tabular Data Protected by Perturbation Methods", Presented at FCSM2009 Conference.

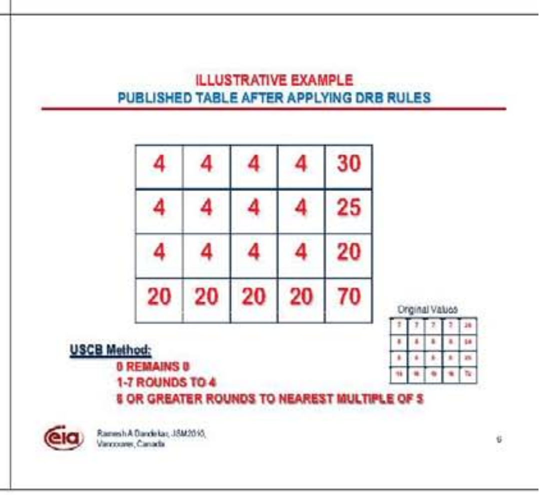
PRESENATION SLIDES











QUALITY (LACK OF) ASPECTS OF PUBLISHED TABLE

- SUM OF INTERNAL COLUMNS =12 => REPORTED 20
- SUM OF INTERNAL ROWS = 16 => REPORTED 30 & 25 & 20
- SUM OF LAST COLUMN = 75 => REPORTED 70
- SUM OF ALL INTERNAL CELLS = 48 => REPORTED 70 (OFF 32%)

4	4	4	4	30
4	4	4	4	25
4	4	4	4	20
20	20	20	20	70

Similar Findings at DOT web site

http://www.fhwa.dot.gov/ctpp/balance.htm

Ramesh A Dandekar, JSM2010, Vancouver, Canada 7

DISCLOSURE POTENTIAL OF INDEPENDENT ROUNDING

Linear Programming (LP) technology allows data users and analysts to recreate bounded additive tabular cell estimates of relatively high precision.

Additive estimates of tabular cells create a *potential* for statistical and exact disclosure of relatively lower count table cells.

MAX OR MIN $C^T x$

SUBJECT TO $A x = b$ LINEAR EQUALITY CONDITIONS

Ramesh A Dandekar, JSM2010, Vancouver, Canada 8

RECREATED TABLE AFTER APPLYING LP-BASED TABULAR ADJUSTMENTS

7	7	7	7	28
7	5	7	7	26
4	6	4	4	18
18	18	18	18	72

IT IS ALWAYS POSSIBLE TO RECREATE ORIGINAL CELL VALUES WITH HIGH PRECISION USING LP TECHNOLOGY

Original Values

7	7	7	7	28
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
18	18	18	18	72

Ramesh A Dandekar, JSM2010, Vancouver, Canada 9

ILLUSTRATIVE EXAMPLE LINEAR PROGRAMMING VARIABLES AND VARIABLE BOUNDS

W01 1-7	W04 1-7	W07 1-7	W10 1-7	W13 28-32
W02 1-7	W05 1-7	W08 1-7	W11 1-7	W14 23-27
W03 1-7	W06 1-7	W09 1-7	W12 1-7	W15 18-22
W16 18-22	W17 18-22	W18 18-22	W19 18-22	W20 68-72

Ramesh A Dandekar, JSM2010, Vancouver, Canada 10

ILLUSTRATIVE EXAMPLE LP_SOLVE FORMAT INPUT

```
+MAX: +1 W01;
+ROW01: +W01 +W04 +W07 +W10 -W13 = 0;
+ROW02: +W02 +W05 +W08 +W11 -W14 = 0;
+ROW03: +W03 +W06 +W09 +W12 -W15 = 0;
+ROW04: +W16 +W17 +W18 +W19 -W20 = 0;
+ROW05: +W01 +W02 +W03 -W16 = 0;
+ROW06: +W04 +W05 +W06 -W17 = 0;
+ROW07: +W07 +W08 +W09 -W18 = 0;
+ROW08: +W10 +W11 +W12 -W19 = 0;
+ROW09: +W13 +W14 +W15 -W20 = 0;
+W01 > 1 ;
+W01 < 7 ;
+W02 > 1 ;
+W02 < 7 ;
+W03 > 1 ;
+W03 < 7 ;
+W04 > 1 ;
+W04 < 7 ;
+W05 > 1 ;
+W05 < 7 ;
+W06 > 1 ;
+W06 < 7 ;
+W07 > 1 ;
+W07 < 7 ;
+W08 > 1 ;
+W08 < 7 ;
+W09 > 1 ;
+W09 < 7 ;
+W10 > 1 ;
+W10 < 7 ;
W11 > 1 ;
W11 < 7 ;
W12 > 1 ;
W12 < 7 ;
W13 > 28 ;
W13 < 32 ;
W14 > 23 ;
W14 < 27 ;
W15 > 18 ;
W15 < 22 ;
W16 > 18 ;
W16 < 22 ;
W17 > 18 ;
W17 < 22 ;
W18 > 18 ;
W18 < 22 ;
W19 > 18 ;
W19 < 22 ;
W20 > 68 ;
W20 < 72 ;
```

W01 1-7	W04 1-7	W07 1-7	W10 1-7	W13 28-32
W02 1-7	W05 1-7	W08 1-7	W11 1-7	W14 23-27
W03 1-7	W06 1-7	W09 1-7	W12 1-7	W15 18-22
W16 18-22	W17 18-22	W18 18-22	W19 18-22	W20 68-72

Ramesh A Dandekar, JSM2010, Vancouver, Canada 11

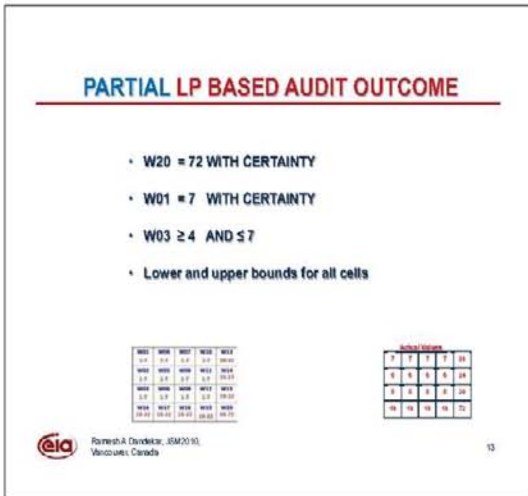
LP_SOLVE OUTPUT

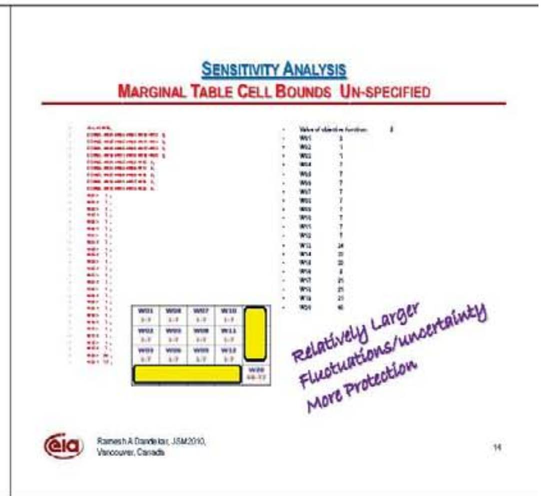
- Value of objective function: 7
- W01 7
- W02 7
- W03 4
- W04 7
- W05 5
- W06 6
- W07 7
- W08 7
- W09 4
- W10 7
- W11 7
- W12 4
- W13 28
- W14 26
- W15 18
- W16 18
- W17 18
- W18 18
- W19 18
- W20 72

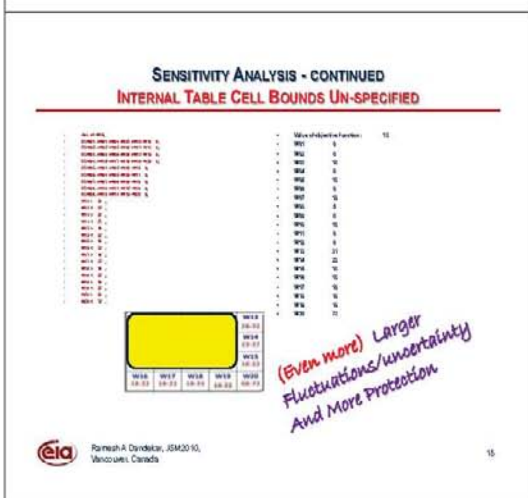
ESTIMATED

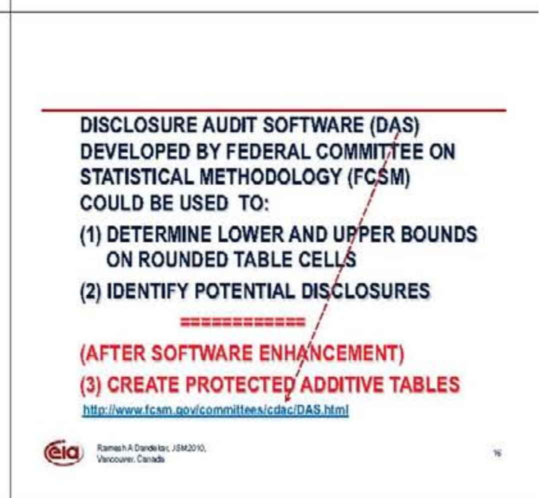
7	7	7	7	28
7	5	7	7	26
4	6	4	4	18
18	18	18	18	72

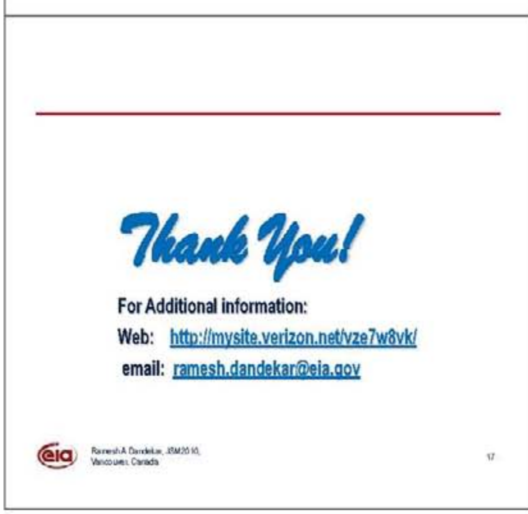
Ramesh A Dandekar, JSM2010, Vancouver, Canada 12


PARTIAL LP BASED AUDIT OUTCOME
• W20 = 72 WITH CERTAINTY
• W01 = 7 WITH CERTAINTY
• W03 ≥ 4 AND ≤ 7
• Lower and upper bounds for all cells
Ramesh A Dandekar, JSM2010,
Vancouver, Canada
13


SENSITIVITY ANALYSIS
MARGINAL TABLE CELL BOUNDS UN-SPECIFIED
Relatively Larger
Fluctuations/uncertainty
More Protection
Ramesh A Dandekar, JSM2010,
Vancouver, Canada
14


SENSITIVITY ANALYSIS - CONTINUED
INTERNAL TABLE CELL BOUNDS UN-SPECIFIED
(Even more) Larger
Fluctuations/uncertainty
And More Protection
Ramesh A Dandekar, JSM2010,
Vancouver, Canada
15


DISCLOSURE AUDIT SOFTWARE (DAS)
DEVELOPED BY FEDERAL COMMITTEE ON
STATISTICAL METHODOLOGY (FCSM)
COULD BE USED TO:
(1) DETERMINE LOWER AND UPPER BOUNDS
ON ROUNDED TABLE CELLS
(2) IDENTIFY POTENTIAL DISCLOSURES
============
(AFTER SOFTWARE ENHANCEMENT)
(3) CREATE PROTECTED ADDITIVE TABLES
http://www.fcsm.gov/committees/cdac/DAS.html
Ramesh A Dandekar, JSM2010,
Vancouver, Canada
16


Thank You!
For Additional information:
Web: http://mysite.verizon.net/vze7w8vk/
email: ramesh.dandekar@eia.gov
Ramesh A Dandekar, JSM2010,
Vancouver, Canada
17