

## **How to Obtain Additive Estimates of Missing Cell Values in Tabular Data Containing Non-Additive Rounded Cells**

**Ramesh A. Dandekar**

**Energy Information Administration (EIA), Washington DC**

**Abstract:** Statistical agencies routinely use complementary cell suppression procedures to protect sensitive cells in tabular data. In addition to the cell suppression procedures, often published tabular cells are also rounded to a certain base value (a multiple of base 10 is common). Deriving additive estimates for suppressed cell values, which corrects for error introduced by the rounding procedure, creates a special challenging situation. By using mathematical programming techniques, in this paper we demonstrate how to simultaneously eliminate the rounding error and obtain additive estimates for suppressed cells so that resultant table is fully populated and has complete additive properties. The technique could be used for other statistical applications such as: a time series of data balancing (e.g. 12 months to annual totals), or a respondent level combined data editing/imputation procedure.

### **Introduction<sup>1</sup>**

Public use of tabular data containing a mix of independently rounded and missing/suppressed data cells often requires pre-processing in order to restore the additive table structure prior to its use for analytical purposes. In this paper we demonstrate how to use linear mathematical optimization techniques to restore the additive table structure by estimating for missing table cell values after correcting for the error resulting from the independent rounding of table cell values. We have used the [LP SOLVE](#) linear programming package available in the public domain to illustrate the procedure required to accomplish this task.

In the first part of the paper we demonstrate a simple method to obtain desired outcome. The method uses one variable to describe each cell in the table structure. The outlined procedure is often used to determine the lower and upper bounds on table cells with missing values. Disclosure Audit Software (DAS) developed by Confidentiality and Data Access Committee (CDAC) under EIA's leadership uses the basic principles described in this paper.

In the second part of the paper we demonstrate a more complex method to obtain maximum likelihood estimates (MLE) of missing and rounded table cell values. The method uses three variables to describe each cell in the table structure. The outlined procedure is typically used to perform a least-absolute-deviation linear regression model on skewed data. For additional information on this method please see [Dandekar2005](#) and [Dandekar2012](#) in the references.

---

<sup>1</sup> This paper is released to encourage discussion and critical comment. The analysis and conclusions expressed here are those of the author(s) and not necessarily those of the U.S. Energy Information Administration (EIA) or the Department of U.S. Energy (DOE).

### Illustrative Example

Table 1 shows hypothetical 4 rows by 5 columns example to illustrate our estimation technique.

**Table 1**

	Column 1	Column 2	Column 3	Column 4	Total Cols
Row 1	350.23	288.03	222.92	332.98	1194.16
Row 2	350.54	374.59	234.71	252.95	1212.79
Row 3	447.82	226.98	390.64	242.11	1307.55
<b>Total Row</b>	<b>1148.59</b>	<b>889.60</b>	<b>848.27</b>	<b>828.04</b>	<b>3714.50</b>

The last column and the last row of the table contain aggregate values from other columns and rows, respectively. The table is completely additive in both dimensions.

For an illustrative purpose, we assume that a decision was made to publish the contents of the Table 1 after rounding the cell values to nearest integer (rounding to base one). After independent rounding of cells, the table appears as shown in Table 2.

**Table 2**

	Column 1	Column 2	Column 3	Column 4	Total Cols
Row 1	350	288	223	333	1194
Row 2	351	375	235	253	1213
Row 3	448	227	391	242	1308
<b>Total Row</b>	<b>1149</b>	<b>890</b>	<b>848</b>	<b>828</b>	<b>3715</b>

In Table 2, the total of the third column and total of second row have rounding errors of one unit (848 vs 849 calculated and 1213 vs 1214 calculated). In many real life table structures

published by statistical agencies, multiple aggregate cells experience non-zero rounding error. The magnitude of the rounding error varies depending upon the number of internal cells over which the aggregate value was obtained. A rounding error of as much as 5 units is rare, but possible. In the case of a multi-dimensional table, the rounding error propagates over all the dimensions, often canceling each.

### Step by Step Procedure for Removing Rounding Error

To eliminate the rounding error by using linear optimization techniques, each cell in Table 2 is assigned a unique variable name. In the Table 3, we make an assignment of variable names (ranging from W01 to W20) to each of the twenty table cells as shown in the red color.

**Table 3**

	Column 1	Column 2	Column 3	Column 4	Total Cols
Row 1	<b>W01</b> 350	<b>W04</b> 288	<b>W07</b> 223	<b>W10</b> 333	<b>W13</b> 1194
Row 2	<b>W02</b> 351	<b>W05</b> 375	<b>W08</b> 235	<b>W11</b> 253	<b>W14</b> 1213
Row 3	<b>W03</b> 448	<b>W06</b> 227	<b>W09</b> 391	<b>W12</b> 242	<b>W15</b> 1308
Total Row	<b>W16</b> 1149	<b>W17</b> 890	<b>W18</b> 848	<b>W19</b> 828	<b>W20</b> 3715

As a next step, all the additive relationships in the table structure need to be captured and clearly identified as an input to mathematical programming solver. In Table 4 we summarize all the nine additive table relations.

**Table 4**

$$\begin{aligned} \text{ROW01: } & + W01 + W04 + W07 + W10 - W13 = 0; \\ \text{ROW02: } & + W02 + W05 + W08 + W11 - W14 = 0; \\ \text{ROW03: } & + W03 + W06 + W09 + W12 - W15 = 0; \\ \text{ROW04: } & + W16 + W17 + W18 + W19 - W20 = 0; \\ \text{ROW05: } & + W01 + W02 + W03 - W16 = 0; \\ \text{ROW06: } & + W04 + W05 + W06 - W17 = 0; \\ \text{ROW07: } & + W07 + W08 + W09 - W18 = 0; \\ \text{ROW08: } & + W10 + W11 + W12 - W19 = 0; \\ \text{ROW09: } & + W13 + W14 + W15 - W20 = 0; \end{aligned}$$

	Column 1	Column 2	Column 3	Column 4	Total Cols
Row 1	<b>W01</b> 350	<b>W04</b> 288	<b>W07</b> 223	<b>W10</b> 333	<b>W13</b> 1194
Row 2	<b>W02</b> 351	<b>W05</b> 375	<b>W08</b> 235	<b>W11</b> 253	<b>W14</b> 1213
Row 3	<b>W03</b> 448	<b>W06</b> 227	<b>W09</b> 391	<b>W12</b> 242	<b>W15</b> 1308
Total Row	<b>W16</b> 1149	<b>W17</b> 890	<b>W18</b> 848	<b>W19</b> 828	<b>W20</b> 3715

The linear programming solver also requires theoretical lower and upper bounds on the value for all variables. In our case the lower and upper bounds for the variables are 0.5

units away from the published independently rounded values in the table. Table 5 summarizes all the lower and upper bounds on the twenty variables.

**Table 5**

W01	<	350	.50;	W11	<	253	.50;
W01	<	349	.500;	W11	<	252	.500;
W02	<	351	.50;	W12	<	242	.50;
W02	<	350	.500;	W12	<	241	.500;
W03	<	448	.50;	W13	<	1194	.50;
W03	<	447	.500;	W13	<	1193	.500;
W04	<	288	.50;	W14	<	1213	.50;
W04	<	287	.500;	W14	<	1212	.500;
W05	<	375	.50;	W15	<	1308	.50;
W05	<	374	.500;	W15	<	1307	.500;
W06	<	227	.500;	W16	<	1149	.500;
W06	<	226	.500;	W16	<	1148	.500;
W07	<	223	.500;	W17	<	890	.500;
W07	<	222	.500;	W17	<	889	.500;
W08	<	235	.500;	W18	<	848	.500;
W08	<	234	.500;	W18	<	847	.500;
W09	<	391	.500;	W19	<	828	.500;
W09	<	390	.500;	W19	<	827	.500;
W10	<	333	.500;	W20	<	3715	.500;
W10	<	332	.500;	W20	<	3714	.500;

In addition to the variable bounds, the linear programming solver also requires an objective function which is a linear function of the form  $\sum a_i w_i$ , where “ $a_i$ ” is weight associated with variable  $w_i$  for  $i = 1$  to 20. These three pieces of information, namely the 1) objective function, 2) additive table relations and 3) lower and upper bounds on the variable serve as an input to the LP\_SOLVE optimizing solver. LP\_SOLVE input file contents are shown in Table 6 below.

**Table 6**

max: + 1 W01;	W08 >	234.50;
ROW01: + W01 + W04 + W07 + W10 - W13 = 0;	W09 >	391.50;
ROW02: + W02 + W05 + W08 + W11 - W14 = 0;	W09 >	390.50;
ROW03: + W03 + W06 + W09 + W12 - W15 = 0;	W10 >	333.50;
ROW04: + W16 + W17 + W18 + W19 - W20 = 0;	W10 >	332.50;
ROW05: + W01 + W02 + W03 - W16 = 0;	W11 >	253.50;
ROW06: + W04 + W05 + W06 - W17 = 0;	W11 >	252.50;
ROW07: + W07 + W08 + W09 - W18 = 0;	W12 >	242.50;
ROW08: + W10 + W11 + W12 - W19 = 0;	W12 >	241.50;
ROW09: + W13 + W14 + W15 - W20 = 0;	W13 >	1194.50;
W01 <	W13 >	1193.50;
W01 >	W14 >	1213.50;
W02 <	W14 >	1212.50;
W02 >	W15 >	1308.50;
W03 <	W15 >	1307.50;
W03 >	W16 >	1149.50;
W04 <	W16 >	1148.50;
W04 >	W17 >	890.50;
W05 <	W17 >	889.50;
W05 >	W18 >	848.50;
W06 <	W18 >	847.50;
W06 >	W19 >	828.50;
W07 <	W19 >	827.50;
W07 >	W20 >	3715.50;
W08 <	W20 >	3714.50;

The input to LP\_SOLVE specifies that objective function for the solution to be a maximum value for the first cell in the table (W01). In reality, multiple options are available for the objective function. It could be individual cell value or aggregates of multiple values. The left half of the Table 7 below shows the output from the LP\_SOLVE program. The right half of the Table 7 is shown to help map the LP\_SOLVE output with relative location estimated cell values in the table.

**Table 7**

	Column 1	Column 2	Column 3	Column 4	Total Cols
Row 1	<b>W01</b> 350	<b>W04</b> 288	<b>W07</b> 223	<b>W10</b> 333	<b>W13</b> 1194
Row 2	<b>W02</b> 351	<b>W05</b> 375	<b>W08</b> 235	<b>W11</b> 253	<b>W14</b> 1213
Row 3	<b>W03</b> 448	<b>W06</b> 227	<b>W09</b> 391	<b>W12</b> 242	<b>W15</b> 1308
Total Row	<b>W16</b> 1149	<b>W17</b> 890	<b>W18</b> 848	<b>W19</b> 828	<b>W20</b> 3715

As you can see, the third column and the second row are additive. This is in addition to additive relations in all other rows and columns.

**Step by Step Procedure to Simultaneously Restore Additivity and Estimate Missing Values**

In situations when some table cell values are either missing or are withheld to protect sensitive table cells, the same linear programming setup could be used. For illustrative purposes we will assume that table cells W02, W03, W08 and W09 values were withheld in Table 8 either to protect sensitive cell values or due to cell value quality concern. The other cells were independently rounded.

**Table 8**

**Table Cells W02, W03, W08, W09 Value Withheld**

	Column 1	Column 2	Column 3	Column 4	Total Cols
Row 1	<b>W01</b> 350	<b>W04</b> 288	<b>W07</b> 223	<b>W10</b> 333	<b>W13</b> 1194
Row 2	<b>W02</b> [withheld]	<b>W05</b> 375	<b>W08</b> [withheld]	<b>W11</b> 253	<b>W14</b> 1213
Row 3	<b>W03</b> [withheld]	<b>W06</b> 227	<b>W09</b> [withheld]	<b>W12</b> 242	<b>W15</b> 1308
Total Row	<b>W16</b> 1149	<b>W17</b> 890	<b>W18</b> 848	<b>W19</b> 828	<b>W20</b> 3715

To estimate the values of withheld cells W02, W03, W08 and W09, we can use the same input format and file that was used to remove the rounding errors. Only missing information in the input file is related to the lower and upper bounds on the four variables. Table 9 shows the input file prepared to solve this problem.



**Table 9**

**Input File for LP\_Solve Program**

```

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 < 350.50;
W01 > 349.50;
W04 < 288.50;
W04 > 287.50;
W05 < 375.50;
W05 > 374.50;
W06 < 227.50;
W06 > 226.50;
W07 < 223.50;
W07 > 222.50;
W10 < 333.50;
W10 > 332.50;
W11 < 253.50;
W11 > 252.50;
W12 < 242.50;
W12 > 241.50;
W13 < 1194.50;
W13 > 1193.50;
W14 < 1213.50;
W14 > 1212.50;
W15 < 1308.50;
W15 > 1307.50;
W16 < 1149.50;
W16 > 1148.50;
W17 < 890.50;
W17 > 889.50;
W18 < 848.50;
W18 > 847.50;
W19 < 828.50;
W19 > 827.50;
W20 < 3715.50;
W20 > 3714.50;
    
```

We have solved the linear programming problem by using two different objective functions, namely for a maximum value of variable W01 and minimum value of variable W01. Table 10 summarizes the linear programming output for these two objective functions. Estimated values for four missing cells are shown in the red font.

**Table 10**

**Additive Missing Cell Estimates**

W01	349.5	min	W01	350.5	max
W02	586.5		W02	0	
W03	213.5		W03	799	
W04	288		W04	288	
W05	374.5		W05	374.5	
W06	227.5		W06	227.5	
W07	222.5		W07	222.5	
W08	0		W08	585.5	
W09	625		W09	39.5	
W10	333.5		W10	333.5	
W11	252.5		W11	252.5	
W12	241.5		W12	241.5	
W13	1193.5		W13	1194.5	
W14	1213.5		W14	1212.5	
W15	1307.5		W15	1307.5	
W16	1149.5		W16	1149.5	
W17	890		W17	890	
W18	847.5		W18	847.5	
W19	827.5		W19	827.5	
W20	3714.5		W20	3714.5	

	Column 1	Column 2	Column 3	Column 4	Total Cols
Row 1	W01	W04	W07	W10	W13
	350	288	223	333	1194
Row 2	W02	W05	W08	W11	W14
	351	375	235	253	1213
Row 3	W03	W06	W09	W12	W15
	448	227	391	242	1308
Total Row	W16	W17	W18	W19	W20
	1149	890	848	828	3715

It is clear from Table 10 that the estimates for missing cells have a “wide” range of values. In practice to improve the quality of estimates for missing cell values, various techniques are used based on the institutional knowledge. The outcome from the institutional knowledge is then transmitted to the linear programming software in terms of bounds (lower and upper) on missing variables. Often point estimates for some of the withheld cells are also used. In Table 11 we have used a guess value for W02 of 293 units, which is an average of the two estimated values of zero and 586 from the linear programming solutions above, to illustrate one of the many possible techniques to narrow the range of additive values.

**Table 11**

**Additive Estimates**  
Using  $W02_{guess} = 293$

```

=====
W01          350.5
W02          293
W03          506
W04          288
W05          374.5
W06          227.5
W07          222.5
W08          292.5
W09          332.5
W10          333.5
W11          252.5
W12          241.5
W13          1194.5
W14          1212.5
W15          1307.5
W16          1149.5
W17           890
W18          847.5
W19          827.5
W20          3714.5
=====
    
```

**Improved Estimates by Using Least Absolute Difference Linear Regression Model**

In the linear programming input setup above, each table cell value is assigned a single variable name. To obtain the linear programming solution close to the centroid of the solution space, each table cell is represented by three variables by using the equation of the form  $W_{Estimate} = W_{XX} + W_{XX\_P} - W_{XX\_M}$  Where  $W_{Estimate}$  is the estimate for the published cell  $W_{XX}$ ; and  $W_{XX\_P}$  and  $W_{XX\_M}$  are respectively the positive and negative corrections required to published cell  $W_{XX}$  to make the entire table additive. For additional information on this method please see [Dandekar2005](#) and [Dandekar2012](#). In the next two tables (Table 12 and 13) we show sample LP\_solve format input for our problem for two different scenarios. In the first scenario there are no withheld cells. Only rounding error needs to be removed. In the second scenario, four table cell values are unknown. The estimates for the four missing cell values are required after removing rounding error.

**Table 12**

**Input File to Remove Rounding Error Only**

```

min: +W01_p +W01_m +W02_p +W02_m +W03_p +W03_m +W04_p +W04_m +W05_p +W05_m +W06_p
+W06_m +W07_p +W07_m +W08_p +W08_m +W09_p +W09_m +W10_p +W10_m +W11_p +W11_m
+W12_p +W12_m +W13_p +W13_m +W14_p +W14_m +W15_p +W15_m +W16_p +W16_m +W17_p
+W17_m +W18_p +W18_m +W19_p +W19_m +W20_p +W20_m ;

ROW00001: +W01_p -W01_m +W04_p -W04_m +W07_p -W07_m +W10_p -W10_m -W14_p +W14_m = 0;
ROW00002: -W02_p +W02_m -W05_p +W05_m -W08_p +W08_m -W11_p +W11_m +W15_p -W15_m = 1;
ROW00003: +W03_p -W03_m +W06_p -W06_m +W09_p -W09_m +W12_p -W12_m -W16_p +W16_m = 0;
ROW00004: +W13_p -W13_m +W17_p -W17_m +W18_p -W18_m +W19_p -W19_m -W20_p +W20_m = 0;
ROW00005: +W01_p -W01_m +W02_p -W02_m +W03_p -W03_m -W17_p +W17_m = 0;
ROW00006: +W04_p -W04_m +W05_p -W05_m +W06_p -W06_m -W18_p +W18_m = 0;
ROW00007: -W07_p +W07_m -W08_p +W08_m -W09_p +W09_m +W19_p -W19_m = 1;
ROW00008: +W10_p -W10_m +W11_p -W11_m +W12_p -W12_m -W13_p +W13_m = 0;
ROW00009: +W14_p -W14_m +W15_p -W15_m +W16_p -W16_m -W20_p +W20_m = 0;

W01_p < 0.5;
W01_m < 0.5;
W02_p < 0.5;
W02_m < 0.5;
W03_p < 0.5;
W03_m < 0.5;
W04_p < 0.5;
W04_m < 0.5;
W05_p < 0.5;
W05_m < 0.5;
W06_p < 0.5;
W06_m < 0.5;
W07_p < 0.5;
W07_m < 0.5;
W08_p < 0.5;
W08_m < 0.5;
W09_p < 0.5;
W09_m < 0.5;
W10_p < 0.5;
W10_m < 0.5;

W11_p < 0.5;
W11_m < 0.5;
W12_p < 0.5;
W12_m < 0.5;
W13_p < 0.5;
W13_m < 0.5;
W14_p < 0.5;
W14_m < 0.5;
W15_p < 0.5;
W15_m < 0.5;
W16_p < 0.5;
W16_m < 0.5;
W17_p < 0.5;
W17_m < 0.5;
W18_p < 0.5;
W18_m < 0.5;
W19_p < 0.5;
W19_m < 0.5;
W20_p < 0.5;
W20_m < 0.5;
    
```

**Table 13**

**Input File to Estimate Missing Cell Values and Rounding Error**

```

min: +W01_p +W01_m +W02 +W03 +W04_p +W04_m +W05_p +W05_m +W06_p +W06_m
      +W07_p +W07_m +W08 +W09 +W10_p +W10_m +W11_p +W11_m +W12_p +W12_m
      +W13_p +W13_m +W14_p +W14_m +W15_p +W15_m +W16_p +W16_m +W17_p
      +W17_m +W18_p +W18_m +W19_p +W19_m +W20_p +W20_m;

ROW000001: +W01_p -W01_m +W04_p -W04_m +W07_p -W07_m +W10_p -W10_m -W14_p +W14_m = 0;
ROW000002: +W02 +W05_p -W05_m +W08 +W11_p -W11_m -W15_p +W15_m = 585;
ROW000003: +W03 +W06_p -W06_m +W09 +W12_p -W12_m -W16_p +W16_m = 839;
ROW000004: +W13_p -W13_m +W17_p -W17_m +W18_p -W18_m +W19_p -W19_m -W20_p +W20_m = 0;
ROW000005: +W01_p -W01_m +W02 +W03 -W17_p +W17_m = 799;
ROW000006: +W04_p -W04_m +W05_p -W05_m +W06_p -W06_m -W18_p +W18_m = 0;
ROW000007: +W07_p -W07_m +W08 +W09 -W19_p +W19_m = 625;
ROW000008: +W10_p -W10_m +W11_p -W11_m +W12_p -W12_m -W13_p +W13_m = 0;
ROW000009: +W14_p -W14_m +W15_p -W15_m +W16_p -W16_m -W20_p +W20_m = 0;

W01_p < 0.5;
W01_m < 0.5;
W04_p < 0.5;
W04_m < 0.5;
W05_p < 0.5;
W05_m < 0.5;
W06_p < 0.5;
W06_m < 0.5;
W07_p < 0.5;
W07_m < 0.5;
W10_p < 0.5;
W10_m < 0.5;
W11_p < 0.5;
W11_m < 0.5;
W12_p < 0.5;
W12_m < 0.5;
W13_p < 0.5;
W13_m < 0.5;
W14_p < 0.5;
W14_m < 0.5;
W15_p < 0.5;
W15_m < 0.5;
W16_p < 0.5;
W16_m < 0.5;
W17_p < 0.5;
W17_m < 0.5;
W18_p < 0.5;
W18_m < 0.5;
W19_p < 0.5;
W19_m < 0.5;
W20_p < 0.5;
W20_m < 0.5;
    
```

The sample summary output from the input files from Table 12 is shown in the Table 14.

**Table 14**

<u>Cell</u>	<u>Before Rounding</u>	<u>Rounded</u>	<u>Additive Estimates</u>	<u>WRT Rounded Difference</u>	<u>WRT Original Difference</u>
W01	350.23	350	350.04167	0.04167	-0.18833
W02	350.54	351	350.875	-0.125	0.335
W03	447.82	448	448.04167	0.04167	0.22167
W04	288.03	288	288.04167	0.04167	0.01167
W05	374.59	375	374.875	-0.125	0.285
W06	226.98	227	227.04167	0.04167	0.06167
W07	222.92	223	222.83333	-0.16667	-0.08667
W08	234.71	235	234.5	-0.5	-0.21
W09	390.64	391	390.83333	-0.16667	0.19333
W10	332.98	333	333.04167	0.04167	0.06167
W11	252.95	253	252.875	-0.125	-0.075
W12	242.11	242	242.04167	0.04167	-0.06833
W13	1194.16	1194	1193.95833	-0.04167	-0.20167
W14	1212.79	1213	1213.125	0.125	0.335
W15	1307.55	1308	1307.95833	-0.04167	0.40833
W16	1148.59	1149	1148.95833	-0.04167	0.36833
W17	889.6	890	889.95833	-0.04167	0.35833
W18	848.27	848	848.16667	0.16667	-0.10333
W19	828.04	828	827.95833	-0.04167	-0.08167
W20	3714.5	3715	3715.04167	0.04167	0.54167



The technique used to solve these two problems could also be used for many other applications of survey operation. One such application is in balancing of time series of data from one collection frequency to another (for example, to make aggregates of three or twelve monthly time series of *preliminary* data to add up to quarterly and annual data collection efforts). In such a scenario internal rows could be used to represent geographic details; such as states, regions and sub-regions and the last row could be used for a national total. The internal columns could be used to represent monthly reported *preliminary* values. The last column could be used to represent either quarterly or yearly reported revised/final data. Bounds on each table cell values will be used to impose percentage tolerance values on the observed preliminary data. Both row and column structure could have imbedded hierarchical table structure. The table of an interest does not have to be restricted to two dimensions and could have multiple hierarchical dimensions and linked structure. The LP input formulation from Table 12 could handle multi-dimensional table of many complexities.

Other potential application areas include combined edit/imputation of respondent level micro data. In recent years many statistical agencies are looking at the feasibility of performing respondent level data editing and imputation for failed edit fields simultaneously. For these efforts integer programming solvers have been proposed and tested. We, however, believe that the combined edit imputation tasks could be better performed by use of linear programming solvers and by following the procedure demonstrated in this paper.


### Conclusion

In this paper we have demonstrated multiple variations of linear optimization techniques that can be used to obtain additive estimates for suppressed cell values in tables containing independently rounded cell values. These techniques have multiple other applications in survey data collection efforts. By using three variables to represent each table cell, the linear programming technique from in this paper uses the least-absolute-difference linear regression analysis technique ([Dandekar2012](#)). The LP techniques proposed in this paper are statistical in nature and therefore, are more appropriate for tabular format data related tasks in statistical surveys.


## References:

- Dandekar R. A. (2001) "[Synthetic Tabular Data: A Better Alternative To Complementary Data Suppression - Original Manuscript Dated December 2001](#)". Also available from CENEX-SDC Project International Conference, PSD2006, Rome, Italy, December 13-15, 2006, Companion CD Proceedings ISBN: 84-690-2100-1. Energy Information Administration, U. S. Department of Energy.
- 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. (2005), "[Complementary Cell Suppression Software Tools for Statistical Disclosure Control - Reality Check](#)", Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality (Geneva, Switzerland, 9-11 November, 2005)
- 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.
- Dandekar Ramesh A. (2010), "[\(In\)Effectiveness of Independent Rounding of Discrete Tabular Data as Statistical Disclosure Control Strategy](#)", [Joint Statistical Meeting 2010](#), Vancouver, Canada, pp 1158-1167, Section on Survey Research Methods.
- Dandekar Ramesh A. (2011), "[Applicability of Basic Separability Principles To Enhance the Operational Efficiency of Synthetic Tabular Data Generation Procedures in Multi Dimensional Table Structures](#)", [Joint Statistical Meeting 2011](#), Miami, Florida, pp 574-585, Section on Survey Research Methods.
- Dandekar Ramesh A. (2012), "[\[In\]Appropriate Use of Statistical Measures in \[the name\] of Balancing Data Quality and Confidentiality of Tabular Format Magnitude Data](#) ", [Joint Statistical Meeting 2012](#), San Diego, California, pp 1842-1856, Section on Survey Research Methods.

## How To Obtain Additive Estimates of Missing Cell Values in Tabular Data Containing Non-Additive Rounded Cells




*Joint Statistical Meetings 2014*  
*Ramesh A. Dandekar, Mathematical Statistician*  
*August 6, 2014, Boston, Massachusetts*

 U.S. Energy Information Administration | *Independent Statistics & Analysis* | www.eia.gov

## Lack of Additivity of Tabular Data


- Independent rounding of table cell values
- Withholding of select table cell values
  1. to protect confidentiality
  2. to reflect quality of information
  3. Insignificant/small quantity

 JSM2014, Boston, Massachusetts | Ramesh A. Dandekar

## Illustrative Example: 4 by 5 Table


**Original Table**

	Column 1	Column 2	Column 3	Column 4	Total Cols
Row 1	350.23	288.03	222.92	332.98	1194.16
Row 2	350.54	374.59	234.71	252.95	1212.79
Row 3	447.82	226.98	390.64	242.11	1307.55
Total Row	1148.59	889.60	848.27	828.04	3714.50

 JSM2014, Boston, Massachusetts | Ramesh A. Dandekar


## Table Values Rounded to Integer Value

	Column 1	Column 2	Column 3	Column 4	Total Cols
Row 1	350	288	223	333	1194
Row 2	351	375	235	253	1213
Row 3	448	227	391	242	1308
Total Row	1149	890	848	828	3715

 JSM2014, Boston, Massachusetts | Ramesh A. Dandekar


## Computational Techniques

- Multiple techniques are possible
- Complexity of linear programming (LP) input setup varies
- Will address simple method with least amount of software coding
- Final outcome differs on how LP problem is set up and which technique is used
- Interior point LP solvers are observed to work better than simplex LP solvers (Dandekar2005)

 JSM2014, Boston, Massachusetts | Ramesh A. Dandekar

## Mathematical Programming Representation of Tabular Data

	Column 1	Column 2	Column 3	Column 4	Total Cols
Row 1	<b>W01</b>	<b>W04</b>	<b>W07</b>	<b>W10</b>	<b>W13</b>
	350	288	223	333	1194
Row 2	<b>W02</b>	<b>W05</b>	<b>W08</b>	<b>W11</b>	<b>W14</b>
	351	375	235	253	1213
Row 3	<b>W03</b>	<b>W06</b>	<b>W09</b>	<b>W12</b>	<b>W15</b>
	448	227	391	242	1308
Total Row	<b>W16</b>	<b>W17</b>	<b>W18</b>	<b>W19</b>	<b>W20</b>
	1149	890	848	828	3715

 JSM2014, Boston, Massachusetts | Ramesh A. Dandekar

### Capturing Additive Relations of Tabular Cell Values

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;

Row	W01	W02	W03	W04	W05	W06	W07	W08	W09	W10	W11	W12	W13	W14	W15	W16	W17	W18	W19	W20
Row1	350	349	351	350	350	350	350	350	350	350	350	350	350	350	350	350	350	350	350	350
Row2	350	349	351	350	350	350	350	350	350	350	350	350	350	350	350	350	350	350	350	350
Row3	351	375	351	375	351	375	351	375	351	375	351	375	351	375	351	375	351	375	351	375
Row4	448	227	448	227	448	227	448	227	448	227	448	227	448	227	448	227	448	227	448	227
Row5	1149	890	1149	890	1149	890	1149	890	1149	890	1149	890	1149	890	1149	890	1149	890	1149	890

©JSM2014, Boston, Massachusetts  
Ramesh A. Dandekar

### Capturing Bounds on Tabular Cell Values

W01 <	350.50;	W11 <	253.50;
W01 >	349.50;	W11 >	252.50;
W02 <	351.50;	W12 <	242.50;
W02 >	350.50;	W12 >	241.50;
W03 <	448.50;	W13 <	1194.50;
W03 >	447.50;	W13 >	1193.50;
W04 <	288.50;	W14 <	1213.50;
W04 >	287.50;	W14 >	1212.50;
W05 <	375.50;	W15 <	1308.50;
W05 >	374.50;	W15 >	1307.50;
W06 <	227.50;	W16 <	1149.50;
W06 >	226.50;	W16 >	1148.50;
W07 <	223.50;	W17 <	890.50;
W07 >	222.50;	W17 >	889.50;
W08 <	235.50;	W18 <	848.50;
W08 >	234.50;	W18 >	847.50;
W09 <	391.50;	W19 <	827.50;
W09 >	390.50;	W19 >	826.50;
W10 <	333.50;	W20 <	3715.50;
W10 >	332.50;	W20 >	3714.50;

©JSM2014, Boston, Massachusetts  
Ramesh A. Dandekar

### Input File for LP\_Solve Program

```

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 < 350.50;
W01 > 349.50;
W02 < 351.50;
W02 > 350.50;
W03 < 448.50;
W03 > 447.50;
W04 < 288.50;
W04 > 287.50;
W05 < 375.50;
W05 > 374.50;
W06 < 227.50;
W06 > 226.50;
W07 < 223.50;
W07 > 222.50;
W08 < 235.50;
W08 > 234.50;
W09 < 391.50;
W09 > 390.50;
W10 < 333.50;
W10 > 332.50;
W11 < 253.50;
W11 > 252.50;
W12 < 242.50;
W12 > 241.50;
W13 < 1194.50;
W13 > 1193.50;
W14 < 1213.50;
W14 > 1212.50;
W15 < 1308.50;
W15 > 1307.50;
W16 < 1149.50;
W16 > 1148.50;
W17 < 890.50;
W17 > 889.50;
W18 < 848.50;
W18 > 847.50;
W19 < 827.50;
W19 > 826.50;
W20 < 3715.50;
W20 > 3714.50;
    
```

©JSM2014, Boston, Massachusetts  
Ramesh A. Dandekar

### Table Cells W02, W03, W08, W09 Value Withheld

	Column 1	Column 2	Column 3	Column 4	Total Cols
Row 1	W01 350	W04 288	W07 223	W10 333	W13 1194
Row 2	W02 [withheld]	W05 375	W08 [withheld]	W11 253	W14 1213
Row 3	W03 [withheld]	W06 227	W09 [withheld]	W12 242	W15 1308
Total Row	W16 1149	W17 890	W18 848	W19 828	W20 3715

©JSM2014, Boston, Massachusetts  
Ramesh A. Dandekar

### Input File for LP\_Solve Program

```

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 < 350.50;
W01 > 349.50;
W02 < 351.50;
W02 > 350.50;
W03 < 448.50;
W03 > 447.50;
W04 < 288.50;
W04 > 287.50;
W05 < 375.50;
W05 > 374.50;
W06 < 227.50;
W06 > 226.50;
W07 < 223.50;
W07 > 222.50;
W08 < 235.50;
W08 > 234.50;
W09 < 391.50;
W09 > 390.50;
W10 < 333.50;
W10 > 332.50;
W11 < 253.50;
W11 > 252.50;
W12 < 242.50;
W12 > 241.50;
W13 < 1194.50;
W13 > 1193.50;
W14 < 1213.50;
W14 > 1212.50;
W15 < 1308.50;
W15 > 1307.50;
W16 < 1149.50;
W16 > 1148.50;
W17 < 890.50;
W17 > 889.50;
W18 < 848.50;
W18 > 847.50;
W19 < 827.50;
W19 > 826.50;
W20 < 3715.50;
W20 > 3714.50;
    
```

©JSM2014, Boston, Massachusetts  
Ramesh A. Dandekar

### Additive Missing Cell Estimates

W01	349.5 min	W01	350.5 max
W02	586.5	W02	0
W03	213.5	W03	799
W04	288	W04	288
W05	374.5	W05	374.5
W06	227.5	W06	227.5
W07	222.5	W07	222.5
W08	0	W08	585.5
W09	625	W09	39.5
W10	333.5	W10	333.5
W11	252.5	W11	252.5
W12	241.5	W12	241.5
W13	1193.5	W13	1194.5
W14	1213.5	W14	1212.5
W15	1307.5	W15	1307.5
W16	1149.5	W16	1149.5
W17	890	W17	890
W18	847.5	W18	847.5
W19	827.5	W19	827.5
W20	3714.5	W20	3714.5

©JSM2014, Boston, Massachusetts  
Ramesh A. Dandekar



### How To Improve Quality/Accuracy of Estimates

- Use of institutional/industry knowledge in combination with appropriate bounds
- By targeting additive estimates towards centroid of solution space (Dandekar2005)
- By using interior point linear programming solvers
- By using three symbolic variables to represent every table cell

### Maximum Likelihood Estimates (MLE)

- Technique demonstrated so far provides additive estimates for table cells
- To obtain MLE, the LP setup is more complex
- Each table cell is represented by using three variables
- Estimate = published + positive adjustment - negative adjustment
- Based on central limit theorem
- Need to use interior point LP solver
- Final solution towards the centroid of the solution space
- "Protecting Sensitive Tabular Data by Complementary Cell Suppression - Myth & Reality", 2005 paper at
- url [http://www.fcsn.gov/05papers/Dandekar\\_IXA.pdf](http://www.fcsn.gov/05papers/Dandekar_IXA.pdf)

### Three Variable LP Model

For each table cell

$$W_{Estimate} = W_{XX} + W_{XX,p} - W_{XX,m}$$

Where

- $W_{XX}$  = Published Value
- $W_{XX,p}$  = Positive Adjustment
- $W_{XX,m}$  = Negative Adjustment

### Input File – Three variable Model – No Missing Cells

```

min: +W01_p +W01_m +W02_p +W02_m +W03_p +W03_m +W04_p +W04_m +W05_p +W05_m +W06_p
+W06_m +W07_p +W07_m +W08_p +W08_m +W09_p +W09_m +W10_p +W10_m +W11_p +W11_m
+W12_p +W12_m +W13_p +W13_m +W14_p +W14_m +W15_p +W15_m +W16_p +W16_m +W17_p
+W17_m +W18_p +W18_m +W19_p +W19_m +W20_p +W20_m ;

ROW00001: +W01_p -W01_m +W04_p -W04_m +W07_p -W07_m +W10_p -W10_m -W14_m +W14_m = 0;
ROW00002: -W02_p +W02_m -W05_p +W05_m -W08_p +W08_m -W11_p +W11_m -W15_m +W15_m = 1;
ROW00003: +W03_p -W03_m +W06_p -W06_m +W09_p -W09_m +W12_p -W12_m -W16_m +W16_m = 0;
ROW00004: +W13_p -W13_m +W17_p -W17_m +W18_p -W18_m +W19_p -W19_m -W20_p +W20_m = 0;
ROW00005: +W01_p -W01_m +W02_p -W02_m +W03_p -W03_m -W17_p +W17_m = 0;
ROW00006: +W04_p -W04_m +W05_p -W05_m +W06_p -W06_m -W18_p +W18_m = 0;
ROW00007: -W07_p +W07_m -W08_p +W08_m -W09_p +W09_m +W19_p -W19_m = 1;
ROW00008: +W10_p -W10_m +W11_p -W11_m +W12_p -W12_m -W13_p +W13_m = 0;
ROW00009: +W14_p -W14_m +W15_p -W15_m +W16_p -W16_m -W20_p +W20_m = 0;

W01_p < 0.5;
W01_m < 0.5;
W02_p < 0.5;
W02_m < 0.5;
W03_p < 0.5;
W03_m < 0.5;
W04_p < 0.5;
W04_m < 0.5;
W05_p < 0.5;
W05_m < 0.5;
W06_p < 0.5;
W06_m < 0.5;
W07_p < 0.5;
W07_m < 0.5;
W08_p < 0.5;
W08_m < 0.5;
W09_p < 0.5;
W09_m < 0.5;
W10_p < 0.5;
W10_m < 0.5;
W11_p < 0.5;
W11_m < 0.5;
W12_p < 0.5;
W12_m < 0.5;
W13_p < 0.5;
W13_m < 0.5;
W14_p < 0.5;
W14_m < 0.5;
W15_p < 0.5;
W15_m < 0.5;
W16_p < 0.5;
W16_m < 0.5;
W17_p < 0.5;
W17_m < 0.5;
W18_p < 0.5;
W18_m < 0.5;
W19_p < 0.5;
W19_m < 0.5;
W20_p < 0.5;
W20_m < 0.5;
    
```

### Interior Point Solver Solution

Cell	Before Rounding	Rounded	Additive Estimates	WRT Rounded Difference	WRT Original Difference
W01	350.23	350	350.04167	0.04167	-0.18833
W02	350.54	351	350.875	-0.125	0.335
W03	447.82	448	448.04167	0.04167	0.22167
W04	288.03	288	288.04167	0.04167	0.01167
W05	374.59	375	374.875	-0.125	0.285
W06	226.98	227	227.04167	0.04167	0.06167
W07	222.92	223	222.83333	-0.16667	-0.08667
W08	234.71	235	234.5	-0.5	-0.21
W09	390.64	391	390.83333	-0.16667	0.19333
W10	332.98	333	333.04167	0.04167	0.06167
W11	252.95	253	252.875	-0.125	-0.075
W12	242.11	242	242.04167	0.04167	-0.06833
W13	1194.16	1194	1193.95833	-0.04167	-0.20167
W14	1212.79	1213	1213.125	0.125	0.335
W15	1307.55	1308	1307.95833	-0.04167	0.40833
W16	1148.59	1149	1148.95833	-0.04167	0.36833
W17	889.6	890	889.95833	-0.04167	0.35833
W18	848.27	848	848.16667	0.16667	-0.10333
W19	828.04	828	827.95833	-0.04167	-0.08167
W20	3714.5	3715	3715.04167	0.04167	0.54167

### Potential Application Areas

- Balancing time series of data from one frequency to other – Making monthly time series of data add up to quarterly and annual time series of data (at state and regional levels to national level).
- Combined edit and imputations
- Estimation of missing values

*Thank You!*

Ramesh A. Dandekar  
 Mathematical Statistician  
 Office of Survey Development and Statistical Integration  
 (202) 386-5845  
 ramesh.dandekar@eia.gov

**Input File – Three variable Model – Four Missing Cells**

```
min: +W01_p +W01_m +W02 +W03 +W04_p +W04_m +W05_p +W05_m +W06_p +W06_m
+W07_p +W07_m +W08 +W09 +W10_p +W10_m +W11_p +W11_m +W12_p +W12_m
+W13_p +W13_m +W14_p +W14_m +W15_p +W15_m +W16_p +W16_m +W17_p
+W17_m +W18_p +W18_m +W19_p +W19_m +W20_p +W20_m

ROW00001: +W01_p -W01_m +W04_p -W04_m +W07_p -W07_m +W10_p -W10_m -W14_p +W14_m = 0;
ROW00002: +W02 +W05_p -W05_m +W08 +W11_p -W11_m -W15_p +W15_m = 585;
ROW00003: +W03 +W06_p -W06_m +W09 +W12_p -W12_m -W16_p +W16_m = 839;
ROW00004: +W13_p -W13_m +W17_p -W17_m +W18_p -W18_m +W19_p -W19_m -W20_p +W20_m = 0;
ROW00005: +W01_p -W01_m +W02 +W03 -W17_p +W17_m = 799;
ROW00006: +W04_p -W04_m +W05_p -W05_m +W06_p -W06_m -W18_p +W18_m = 0;
ROW00007: +W07_p -W07_m +W08 +W09 -W19_p +W19_m = 625;
ROW00008: +W10_p -W10_m +W11_p -W11_m +W12_p -W12_m -W13_p +W13_m = 0;
ROW00009: +W14_p -W14_m +W15_p -W15_m +W16_p -W16_m -W20_p +W20_m = 0;

W01_p < 0.5; W13_p < 0.5;
W01_m < 0.5; W13_m < 0.5;
W04_p < 0.5; W14_p < 0.5;
W04_m < 0.5; W14_m < 0.5;
W05_p < 0.5; W15_p < 0.5;
W05_m < 0.5; W15_m < 0.5;
W06_p < 0.5; W16_p < 0.5;
W06_m < 0.5; W16_m < 0.5;
W07_p < 0.5; W17_p < 0.5;
W07_m < 0.5; W17_m < 0.5;
W10_p < 0.5; W18_p < 0.5;
W10_m < 0.5; W18_m < 0.5;
W11_p < 0.5; W19_p < 0.5;
W11_m < 0.5; W19_m < 0.5;
W12_p < 0.5; W20_p < 0.5;
W12_m < 0.5; W20_m < 0.5;
```

Least Absolute Difference  
 Linear Regression Estimates  
 Aka L1 norm Regression

**Interior Point Solver Solution**

Cell	Original	Additive Estimates	Difference	Missing Value Estimate
W01	350	350.04167	0.04167	350
W02	351	350.875	-0.125	328.85888 w
W03	448	448.04167	0.04167	470.14112 w
W04	288	288.04167	0.04167	288
W05	375	374.875	-0.125	375
W06	227	227.04167	0.04167	227
W07	223	222.83333	-0.16667	223
W08	235	234.5	-0.5	256.14112 w
W09	391	390.83333	-0.16667	368.85888 w
W10	333	333.04167	0.04167	333
W11	253	252.875	-0.125	253
W12	242	242.04167	0.04167	242
W13	1194	1193.95833	-0.04167	1194
W14	1213	1213.125	0.125	1213
W15	1308	1307.95833	-0.04167	1308
W16	1149	1148.95833	-0.04167	1149
W17	890	889.95833	-0.04167	890
W18	848	848.16667	0.16667	848
W19	828	827.95833	-0.04167	828
W20	3715	3715.04167	0.04167	3715