Proceedings of Statistics Canada Symposium 2001
Achieving Data Quality in a Statistical Agency: a Methodological Perspective

DISCLOSURE AUDITING IN ROUNDED TABLES

Nancy J. Kirkendall, Ruey-Pyng Lu, Mark A. Schipper,[1] Stephen F. Roehrig[2]

ABSTRACT

It is customary for statistical agencies to audit tables containing suppressed cells to ensure that there is sufficient protection against inadvertent disclosure of sensitive information. If the table contains rounded values, this fact may be ignored by the audit procedure. This oversight can result in over-protection, reducing the utility of the published data. In this paper, we provide a correct auditing formulation, and give examples of over-protection.

KEY WORDS: Cell Suppression, Rounding, Confidentiality.

1. INTRODUCTION

Statistical agencies routinely publish aggregate tables containing values rounded to the closest multiple of some base amount. For example, the U.S. Bureau of Economic Analysis publishes an annual survey of U.S. direct investments abroad, comprising tables that give dollar figures for industry by country, rounded to the nearest $1 million. Often, some cells in the tabulation are considered sensitive, and so are suppressed in the actual publication. To prevent the inference of a suppressed value from the published marginal totals, additional cells called complementary suppressions may also be suppressed. The procedure for determining the complementary suppressions is usually heuristic, since any optimal procedure is known to be NP-complete for tables of dimension three or higher (Kelly, 1992). Therefore, it is of great interest to check the results of the heuristic to ensure the required level of protection actually obtains.

The procedure usually used for auditing is linear programming (Sande, 1999 and Zayatz 1993). Although for two-dimensional tables there exist fast network procedures for disclosure auditing (Gusfield, 1988), linear programming (LP) is customary, since it can also be used for higher-dimensional tables where network models break down. The conventional formulation of the LP problem, in use for at least a decade by many U.S. and foreign statistical agencies, is flawed for rounded tables. In this paper, we point out the error in the conventional analysis, present a corrected formulation, and provide results that give an indication of the level of over-protection (and hence loss of data utility) that may result from the flawed auditing procedure.

2. LP FORMULATION

Suppose that Table 1 gives the true data, in decimal form, collected by a statistical agency. In this table, row 0 contains the marginal sums of rows 1-4, and column 10 has the marginal totals of columns 101-104.

	10	101	102	103	104
0	170.2	39.5	58.8	44.1	27.8
1	35.0	14.2	15.3	2.2	3.3
2	41.3	5.4	10.2	10.3	15.4
3	47.1	7.8	20.0	15.0	4.3
4	46.8	12.1	13.3	16.6	4.8

Table 1: Original Table of Unrounded Data

[1] Nancy J. Kirkendall, Ruey-Pyng Lu, Mark A. Schipper, Statistical Methods Group, Energy Information Agency, 1000 Independence Ave., SW, Washington, DC USA 20585.
[2] Stephen F. Roehrig, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA USA 15213

The data agency wishes to publish this table, after rounding its entries to whole numbers, and after applying disclosure protection in the form of cell suppression. There are several rounding methods in common use. The simplest is independent rounding, in which each cell is rounded, in isolation of the others, to the closest multiple of the rounding base. If independent rounding is applied to Table 1,a non-additive table (i.e., one in which a row or column of rounded values doesn't sum to the rounded marginal value) results. In spite of this possibility, independent rounding is often used because of its simplicity. Various means exist for *controlled rounding*, methods that are more complex but do produce additive tables (e.g., Cox 1987). Our analysis applies, with minor variations, to tables subjected to either form of rounding. To audit a non-additive, independently rounded table using linear programming, an additional step is required, which is to produce a best-fit, additive, not-necessarily-integer table. These details are not discussed here.

Suppose that Table 2 is the result of the rounding process, and the shaded cells shown have been chosen for suppression. The linear programming audit procedure consists in taking the suppressed cells as variables, then maximizing and minimizing their possible values, one-by-one, subject to constraints that arise from the published interior cells and marginal totals.

	10	101	102	103	104
0	170.0	40.0	58.0	44.0	28.0
1	35.0	14.0	15.0	2.0	4.0
2	41.0	6.0	10.0	10.0	15.0
3	47.0	8.0	20.0	15.0	4.0
4	47.0	12.0	13.0	17.0	5.0

Table 2: Rounded Table, Suppressions Shaded

The constraint set for Table 2 is

$$x_{(1,103)} + x_{(1,104)} = 6$$

$$x_{(3,103)} + x_{(3,104)} = 19$$

$$x_{(1,103)} + x_{(3,103)} = 17$$

$$x_{(1,104)} + x_{(3,104)} = 8$$

Auditing the protection for the suppressed cells, using any LP solver, results in the (lower, upper) bounds in Table 3.

	103	104
1	(0, 6)	(0, 6)
3	(11, 17)	(2, 8)

Table 3: Bounds on Suppressed Cells

This conventional approach to disclosure auditing assumes that the rounded values are exact. Despite the fact that data users, and data intruders, are normally fully advised that the table entries are rounded, this simplifying assumption is very commonly made. This oversight can result in misleading evaluations of data protection. Depending on whether suppression is applied to the rounded or unrounded data, over- or under-suppression can result. In the case of under-suppression, the statistical agency may have failed to protect the confidentiality of its respondents; for over-suppression, data users suffer the disutility of unnecessary suppressions.

2.1 Energy Information Administration Example

Protecting respondent confidentiality is a high priority in the federal statistical community. Indeed, much of the data collected by statistical agencies is collected under an explicit promise of confidentiality; that is, that information will not be released in a way that allows users of information to determine the data values for individual respondents or a grouping of respondents. Data collected under such a promise are commonly made available to data users in the form of aggregated (tabular) data summaries. Because such releases have a risk of disclosing confidential data, agencies employ a variety of safeguards in an attempt to avoid disclosures. Indeed, agencies give high priority to assessing disclosure risks and taking necessary steps (data suppression, error inoculation, and so on) to minimize (ideally, eliminate) such risks. Although rare, there have been instances where official data releases have allowed confidential data to be identifiable. More likely, although an untested theory, there have been instances where official data releases have seemingly breached confidentiality protection rules because of a simplifying assumption, such as assuming rounded summary values are exact.

We give an example from the Energy Information Administration (EIA) report "*Manufacturing Consumption of Energy 1991,*" which is funded and disseminated by EIA but collected, edited and tabulated by the US Bureau of the Census. Complexities of tabular summaries found in this EIA report vary. We have chosen a table that offers some complexity because of its disclosure pattern, but tables that are more complex exist in this EIA report. Table A28 presents statistics on US manufacturing expenditures (in million dollars) for *Purchased Energy* by six establishment size categories and four Census Regions, yielding a 3-D table structure. Of the 215 cell values in Table A28, 87 were suppressed from publication.

For this paper, we reduce the display of this table to those cells of interest. Table 4 shows *Expenditures for Purchased Energy* for *Distillate Fuel Oil* for each *Census Region* from Table A28. An examination of this reduced table reveals how a user may make a misleading evaluation when rounded cell values are assumed exact.

	Total	Northeast	Midwest	South	West
Total	800	248	152	276	124
Under 20	351	132	W	W	54
20-49	154	40	W	W	W
50-99	76	21	13	26	16
100-249	71	22	12	25	12
250-499	53	W	8	27	W
500 +	95	W	31	35	W

Table 4: Rounded Table of Expenditures, Table A28

First, re-order the Northeast column of Table 4 to obtain Table 5. Re-ordering simply allows for a better display of the suppression pattern found in the reduced table.

	Total	Midwest	South	West	Northeast
Total	800	152	276	124	248
Under 20	351	W	W	54	132
20-49	154	W	W	W	40
50-99	76	13	26	16	21
100-249	71	12	25	12	22
250-499	53	8	27	W	W
500 +	95	31	35	W	W

Table 5: Rounded Table of Expenditures, Re-Ordered

In Table 5, we have divided withheld values into three groupings: two are bordered and one is bordered and gray-shaded. These groupings are important. Because of the overlap in suppression patterns – under an assumption that cells are exact (i.e., *not* rounded) – one may derive an integer value for the shaded cell (*Distillate Fuel Oil* expenditures in the *West* Census Region for the *20-49* establishment size category). Further, this is a familiar example in which every row and column has at least two suppressions, yet protection is incomplete, at least for tables where rounded cell values are considered exact.

Let's derive that *integer, gray-shaded* cell of Table 5. Using the linear relationships of the marginal totals, we calculate the grouped withheld values for *Under 20* and *20-49* as 165 (351 minus 186) and 114 (154 minus 40), respectively. When combined, these values sum to 279. Similarly, withheld values for the *Midwest* and *South* are 88 (152 minus 64) and 163 (276 minus 113), respectively. Again, when summed, these withheld values equal 251. Taken separately, these values of 279 and 251 have the intended result of minimizing disclosure risks. However, these same values become valuable to a data user – or data intruder

– because we may easily derive, by subtraction, a seemingly *exact* disclosure of the withheld value for *Distillate Fuel Oil* expenditures in the *West* Census Region for *20-49* establishment size category. That value is 28 (279 minus 251).

Surprisingly, for our shaded *Distillate* cell, we obtain the same value (28) using the other remaining withheld values. That is, we may calculate a value for the *West* (42) and *Northeast* (33) Census Regions and the *249-500* (18) and *500+* (29) establishment size categories and subtract these summed values to obtain 28 for the shaded *Distillate Fuel Oil* value.

Not surprisingly, a data user or data intruder might find this information useful. Moreover, and more importantly, statistical agencies find this information harmful to their mission and perceived ability to protect respondents from unacceptable disclosure risks. Fortunately, both results may be false because this table has an inherent rounding process; that is, the simplifying assumption that rounded cells values are exact does not hold for Table A28.

Auditing the disclosure protection for the suppressed cells, using an LP formulation that recognizes the rounding process, results in the (lower, upper) bounds in Table 6.

	Northeast	Midwest	South	West
Under 20		(0,84)	(79.5,165)	
20-49		(2.5,91)	(0,85.5)	(21,34)
250-499	(0,19.5)			(0,19.5)
500 +	(11,30.5)			(0,19.5)

Table 6: Bounds on Expenditures Table

This agency example demonstrates the impact of oversimplifying disclosure auditing. Specifically, we show in Table 6 that a seemingly *exact* disclosure – 28 for the *Distillate Fuel Oil* expenditures in the *West* Census Region – is actually a misleading evaluation because the rounding process provides lower and upper bounds of 21 and 34, respectively. Because these data are confidential, it is impossible to ascertain the direct impact of these feasible bounds on privacy; however, the error of taking rounded values as exact is empirically revealed. Indeed, this example provides direct evidence of the potential errors – misleading evaluations on disclosure controls – introduced when disclosure audits are inaccurately formulated.

2.1 Corrected Formulation

The solution to these problems is simple and obvious. All that needs to be done is to include additional variables and constraints for the rounded cell values, explicitly specifying the rounding bounds. This introduces additional complexity into the linear programming problem, but assures that the disclosure auditing is done correctly.

Constraints appropriate for Table 2, row 1, for example, are the following.

$$x_{(1,101)} + x_{(1,102)} + x_{(1,103)} + x_{(1,104)} = x_{(1,10)}$$

$$13.5 \le x_{(1,101)} \le 14.5$$

$$14.5 \le x_{(1,102)} \le 15.5$$

$$34.5 \le x_{(1,10)} \le 35.5$$

On solving the corrected formulation, the (lower, upper) bounds for the suppressed cells are those in Table 7.

	103	104
1	(0, 7.5)	(0, 7.5)
3	(8, 18.5)	(0, 9.5)

Table 7: Correct Bounds

3. ERROR ANALYSIS

We can give an analytic account of how the bounds for a suppressed cell expand when rounded values are taken as exact values. We do this for two-dimensional tables only, because there is a convenient network representation for 2-D tables that doesn't apply to 3-D and higher tables. Starting with the simple 4×4 table in Table 2, we follow Gusfield (Gusfield,1988) and draw a network corresponding to the suppressed cells (Figure 1).

In the network, nodes represent rows (on the left) and columns (on the right). Pairs of arcs join the nodes, one pair for each suppressed cell. In this graph, flows along arcs represent potential values that the suppressed cells may assume. The upper arc in each pair holds the current value a suppressed cell. The lower arc holds a value equal to the largest row sum of the suppressed cell values. The purpose of the lower arcs is to record the maximum flow that could occur on any arc in the network.

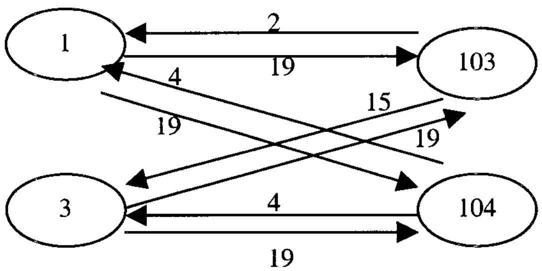

Figure 1: Graph of Suppressed Cells

To determine the upper and lower bounds for any suppressed cell, linear programming can be used, but Gusfield's network approach provides insight into the effects of rounding, at least in 2-D tables. As is usual in network flow models, the maximum flow from one node to another along a specific directed path (i.e., a succession of consecutive arcs) is equal to the minimum of the potential flows along the component arcs in the path. To find the maximum and minimum for a suppressed cell, first define M to be the largest of the rowsums of suppressed cells (19 in our example). This is the value associated with the lower arc in each pair in the graph. Also, let T(i,j) be the current table value for the suppressed cell in row i and column j. This is the value of the upper arc in each pair. Finally, define F(i,j) to be the maximum flow from node i to node j, taken as the sum over *all* directed paths joining i to j.

The following result is central.

Theorem 1: *(Gusfield 1988, Theorem 5) The upper bound for suppressed cell (i*, j*) is equal to P(j*, i*), the total flow from j* to i* (note the reversed order of the nodes here). The lower bound for cell (i*, j*) equals max[0, T(i*, j*) – P(i*, j*) + M].*

For example, the upper bound for the suppressed cell (3, 103) is the sum 15+2=17 of the direct flow 103→3 and the indirect flow 103→1→104→3. The flow P(3, 103) is the sum of the direct path 1→103 plus the indirect path 3→104→1→103, for a total of 19+4=23. Thus the lower bound for the suppressed cell (3, 103) is max[0, 15 – 23 + 19] = 11.

We can now consider what happens when we know that certain unsuppressed cells are rounded. We first consider the case when all the internal cells are rounded, but marginal totals are still considered exact. We think of a rounded cell as being "partially suppressed," that is, its exact value isn't known but bounds on it are known. For the standard case of a cell rounded to a whole number, these bounds are the published cell value ± 1/2.

We include the rounded cells in the network, with each upper arc from a rounded cell holding the value 1/2. This accounts for the ambiguity in knowing its value. This new network will have, for our example problem, four row nodes and four column nodes, with each row node connected by an arc pair to each column node. With this done, the upper and lower bounds on suppressed cells are calculated much as in Gusfield's theorem. Note that there will be additional paths that need to be considered in each calculation, these new paths going along an arc for a rounded cell.

Theorem 2: *Bounds for suppressed cells in rounded tables are computed using Gusfield's algorithm on the expanded graph.*

As an example, we calculate the upper and lower bounds for the suppressed cell (3, 103). For the upper bound, we sum over the original paths 103→3 and 103→1→104→3, giving the value of 17 we obtained previously. In addition, however, we include the paths 103→2→104→3 and 103→4→104→3, both with flow 1/2. Adding up all these flows, we find the true upper bound for cell (1, 103) to be 18 rather than the value of 17 obtained before. (When determining the flows, one needs to be careful that two flows do not "re-use" an arc in such a way that the total flow over all paths carried by that arc exceeds its capacity. Thus having chosen path 103→2→104→3, one could not also use 103→2→102→3, since the capacity of the arc 103→2 is restricted to 1/2. Also, no cell value is permitted to become negative as a result of multiple flows.)

For the lower bound on cell (3, 103), the total flow P(3, 103) can be constructed using the following paths: 3→103, 3→104→2→103, 3→104→4→103, 3→101→1→103, and 3→102→1→103. These total 25, so the lower bound is max[0, 15 – 25 + 19] = 9.

Note that in both of these examples, the result of following each path amounts to an exchange of values between cells. These exchanges are continued until no more are possible, because of bound or non-negativity restrictions. In general, there may be many ways to maximize or minimize a suppressed cell, resulting in the same maximum or minimum value, but different configurations of the other table cells. Moving from one optimal configuration to another is a move known, in LP language, as a degenerate pivot.

It's now possible to see what would happen if the table were larger. For example, if the table were 10×10, but still had the simple 2×2 suppression pattern, each of the upper bounds could increase by up to 1/2(10 – 2) = 4 over the same bound computed under the assumption that the rounded cells are exact. Lower bounds can potentially expand by twice this much. This has clear ramifications for choosing, and auditing, suppression patterns.

Suppose, in the example table, the cell (3, 103) was a primary suppression. The other three suppressions might easily be chosen as complements, because their sum is a minimum amongst all available rectangular four-cell suppression patterns. If a disclosure audit were performed under the assumption that the rounded cells are exact, then the audit would find bounds of [11, 17] around the "exact" value of 15. Since the upper bound is only 13% larger than the "exact" value, this might prompt the agency to search for a different suppression pattern, or perhaps more likely, augment the existing pattern with additional suppressions. If the agency rule were that secondary suppressions should protect to within 20%, that rule would *in fact* be satisfied by the current pattern. Thus, the assumption that the rounded values are exact may easily result in unnecessary suppressions, reducing the value of the published data product.

The analysis presented so far can be extended to the case where marginal totals are also rounded. To do this, the network must be expanded to include nodes representing the marginal totals. The resulting network is precisely the same as that used by, e.g., Cox (Cox 1995), and is shown for a rounded 4×4 table in Figure 2.

Extending the theorem of Gusfield to this network is also straightforward. The net effect is to include additional ambiguity into the calculation of upper and lower cell bounds. For the simple "rectangular" suppression example we have been considering, upper

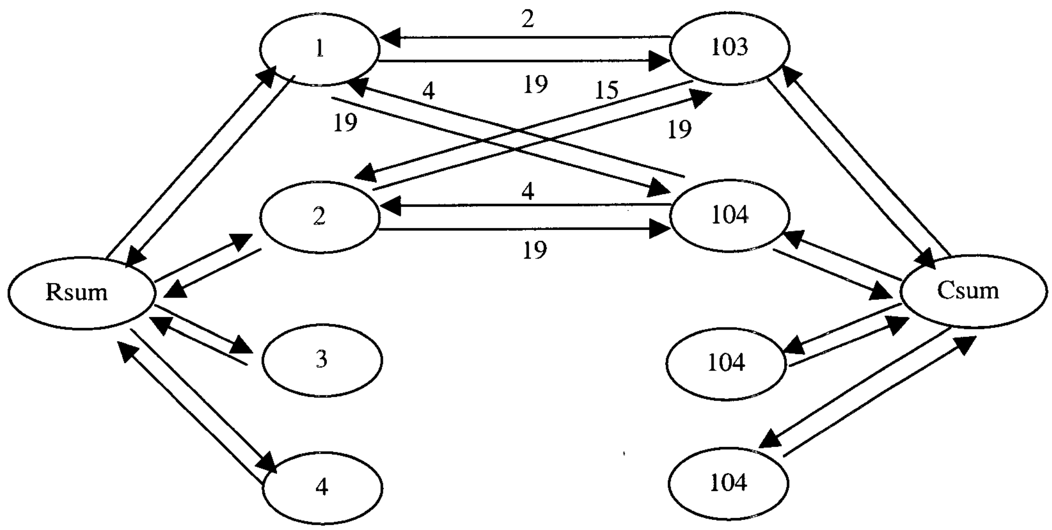

Figure 2: Network Including Marginal Totals

bounds on each suppressed cell are increased by an additional 1/2, while each lower bound decreases by 1.0 (unless zero is reached).

For tables with larger numbers of suppressions, the effect of rounding is harder to analyze, because of the potential for interaction among cells. As an extreme case, Table 8 gives another version of our example with additional suppressions (indicated by shading).

	10	**101**	**102**	**103**	**104**
0	170	40	**59**	44	**28**
1	**35**	14	15	**2**	3
2	41	**5**	10	10	**15**
3	**47**	8	20	**16**	4
4	47	**12**	**13**	**16**	5

Table 8: Table with Additional Suppressions

In this example, if we assume that the cell values are exact, cell (4, 103) has identical upper and lower bounds, constituting an *exact* disclosure. This is an example of the familiar pattern in which every row and column has either zero, or at least two, suppressions, yet protection is incomplete. However, if we recognize that the cell values are rounded, the true bounds for this cell are [10.5, 21.5], obviously a substantial difference.

For three-dimensional tables (and higher), there is no longer a direct network representation of the table. Thus, it is again difficult to analyze the effects of rounding. Empirically, however, results similar to those demonstrated above make it clear that assuming rounded cells are exact can result in over-suppression.

Bureau of Economic Analysis Example

The Bureau of Economic Analysis offers a larger table for examining disclosure audits. Table 9 shows a portion of data from Table 28 found in the BEA report, "*Direct Investment Abroad 1991.*" Using the same technique of isolating cells, we may derive an integer value for two cells: *Tobacco* for *Canada* and *Africa*.

	Total	Canada	Europe	Latin America	Africa	Middle East	Pacific	International
Other Mfg.	126,840	19,992	77,307	15,429	652	69	13,334	57
Tobacco	14,497	d	9,147	3,085	d	0	725	0
Textile	5,782	466	4,221	436	0	0	659	0
Lumber	3,392	686	2,071	350	0	0	284	1
Paper	30,088	7,794	17,424	2,721	d	d	2,046	0
Printing	5,726	1,141	3,654	215	33	0	683	0
Rubber	7,371	985	3,755	1,702	d	0	824	d
Plastics	3,495	430	2,423	161	18	0	463	0
Glass	3,640	d	2,351	607	0	0	d	0
Stone	5,327	2,677	1,832	397	d	0	358	d
Instruments	44,485	3,889	28,498	5,538	d	d	6,409	0
Other	3,037	d	1,931	217	0	0	d	0

Table 9: Rounded Table of Investments, Table 28

Re-ordering the rows and columns of Table 9 provides a better display of the suppression pattern and results in the following:

	Total	Canada	Africa	Middle East	Pacific	International	Europe	Latin America
Other Mfg.	126,840	19,992	652	69	13,334	57	77,307	15,429
Other	3,037	d	0	0	d	0	1,931	217
Glass	3,640	d	0	0	d	0	2,351	607
Tobacco	14,497	d	d	0	725	0	9,147	3,085
Paper	30,088	7,794	d	d	2,046	0	17,424	2,721
Instruments	44,485	3,889	d	d	6,409	0	28,498	5,538
Rubber	7,371	985	[redacted]	0	824	[redacted]	3,755	1,702
Stone	5,327	2,677	[redacted]	0	358	[redacted]	1,832	397
Plastics	3,495	430	18	0	463	0	2,423	161
Printing	5,726	1,141	33	0	683	0	3,654	215
Textile	5,782	466	0	0	659	0	4,221	436
Lumber	3,392	686	0	0	284	1	2,071	350

Table 10: Re-Ordered, Rounded Table of Investments

In Table 10, we have formed four groupings of withheld ("d") cells: *Other* and *Glass*; *Tobacco*; *Paper* and *Instruments*; and, *Rubber* and *Stone*. Using these combinations, if rounded cells values are considered exact, we may determine integer values for the each of gray-shaded cells: *Tobacco* for *Canada* and *Africa*. Here's how.

By subtraction, we calculate the withheld values for *Canada*, *Africa*, *Middle East* and *International* as 1,924; 601; 69; and 56 respectively. Similarly, "d" values for *Rubber* and *Stone* are 105 and 63; and, when combined, they equal 168 (105 plus 63). From the combined withheld values for *International* (56) and *Rubber* and *Stone* (168), we may isolate the suppressed value for *Rubber* and *Stone* for *Africa* as 112 (168 minus 56). Likewise, by subtraction, we may isolate the suppressed value for *Paper* and *Instruments* for *Africa* as 185 (254 minus 69).

Finally, the value for *Tobacco* for *Africa* may be given as 304 (601 minus 112 minus 185). Further, the value for *Tobacco* for *Canada* is 1,236 (1540 minus 304). These values, however, are misleading because cells values are not exact; rather they are rounded to the nearest $1 million.

Applying a correctly formulated set of LP constraints (i.e., one that recognizes a table's rounding process) to any linear optimizer, we may calculate the (lower, upper) bounds for the suppressed cells. We display these audit results in Table 11.

	Total	Canada	Europe	Latin America	Africa	Middle East	Pacific	International
Other Mfg.								
Tobacco		(1,223.5, 1,248.5)			(291, 317)			
Textile								
Lumber								
Paper					(31, 105.5)	(0, 69.5)		
Printing								
Rubber					(45.5, 107.5)			(0, 57)
Plastics								
Glass		(0, 683.5)					(0, 683,5)	
Stone					(3.5, 65.5)			(0, 57)
Instruments					(79, 153.5)	(0, 69.5)		
Other Other		(0, 696)					(194.5, 888)	

Table 11: Correct Bounds of Investments

Table 11 reveals two insights. First, as already mentioned, re-formulated LP constraints provide direct evidence of the error associated with simplifying assumptions. Second, Table 11 reveals that large tables may have wider (upper and lower) bounds because of the increased contribution of the rounding process to disclosure coverage.

4. CONCLUSIONS

When publishing tables incorporating cell suppressions, a data disseminator tries to strike a balance between privacy protection and data utility. Complementary suppressions are generally chosen to achieve a desired interval of protection while minimizing their number. In the case of rounded tables, a common error is to take the rounded cells as exact, often resulting in over-protection. In this paper, we have analyzed the consequences of this error, and provided a simple repair.

REFERENCES

Cox, L. H. (1995),"Network models for complementary cell suppression", *Journal of the American Statistical Association,* 90, pp. 1453-1462.

Cox, L. H. (1987), "A constructive procedure for unbiased controlled rounding", *Journal of the American Statistical Association,* 82, pp 38-45.

Gusfield, D. (1988, "A Graph Theoretic Approach to Statistical Data Security", *Siam Journal of Computing* 17, pp. 552-571.

Kelly, J., Golden, B., and Assad, A. (1992), "Cell suppression: disclosure protection for sensitive tabular data", *NETWORKS,* 22, pp. 397-417.

Sande, G. (1999), "Structure of the ACS Automated Cell Suppression System", Working Paper # 9, presented at the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Thessaloniki, Greece.

Zayatz, L. (1993), "Using linear programming methodology for disclosure avoidance purposes", *Proceedings of the International Seminar on Statistical Confidentiality.* EUROSTAT, Luxembourg, pp. 341-351.