

Working Paper No. 40
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 ECE/Eurostat work session on statistical data confidentiality
(Luxembourg, 7-9 April 2003)

Topic (vi): Software tools for statistical disclosure control

**COST-EFFECTIVE IMPLEMENTATION OF SYNTHETIC TABULATION
(A.K.A. CONTROLLED TABULAR ADJUSTMENT)
IN LEGACY AND STATE OF THE ART STATISTICAL DATA PUBLICATION SYSTEMS**

Contributed paper

Submitted by the Department of Energy, United States¹

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

**Cost Effective Implementation of Synthetic Tabulation
(a.k.a. Controlled Tabular Adjustment)
In Legacy and State of the Art Statistical Data Publication Systems**

Ramesh A. Dandekar
(Ramesh.Dandekar@eia.doe.gov)
Statistics and Methods Group, EI-70
U. S. Department of Energy
Washington DC 20585-0670

ABSTRACT

Recently, Dandekar/Cox (2002) have proposed using controlled tabular adjustments (CTA) to prevent statistical disclosure of sensitive information in tabular data systems. Irrespective of the relative ease with which CTA produces synthetic tables, the cumulative conversion cost associated with a large number of legacy data systems could create a monumental financial burden for the National Statistical Offices (NSO). To minimize the cumulative financial impact on NSO, we further simplify the basic CTA implementation procedure. We also provide the rationale for simplification, followed by a step-by-step implementation of our proposed procedure. Illustrative examples using complex tabular test cases are used to demonstrate the effectiveness of the modified procedure.

INTRODUCTION

As an alternative to the cell suppression method, Dandekar/Cox (2002) proposes applying controlled tabular cell adjustments to tables containing sensitive cells. The major thrust of the CTA implementation procedure is based on the fact that “relatively minor” changes to “statistical” table cell values do not degrade the quality of the statistical information. The quality of the statistical information is typically assessed in terms of percent error (or percent accuracy) of published values. So long as the percent error is statistically insignificant, the usefulness of the altered cell values remains undiminished.

Currently many NSO use the cell suppression method to protect tabular data from statistical disclosure. For these NSO, converting from cell suppression to the CTA method could be an undertaking of considerable cost. We, therefore, propose a very simplistic modification to the Dandekar/Cox (2002) CTA method. The modified procedure could be implemented with minimal changes to existing software languages and code currently used by NSO to generate tabular data.

MATHEMATICAL AND STATISTICAL BASIS

In any table structure, irrespective of imbedded complexities arising from hierarchical/linked properties, the marginal cell value is generally larger than the corresponding internal tabular cell values. There are only two exceptions to this rule. One exception is when there is only one internal cell corresponding to the given marginal cell. Another exception occurs when some internal table cell values are negative.

The Dandekar/Cox (2002) method adjusts sensitive cell values away from their true values to reduce statistical disclosure, and thereafter minimally adjusts non-sensitive cell values to restore original additive properties of the table structure. As mentioned earlier, minimizing percent error to non-sensitive cells gains importance in the decision process to retain both validity and quality of statistical information. Hence the selection of related marginal cells to counter-balance the adjustment made to internal sensitive cells offers maximum benefit in terms of minimal percent error.

STEP-BY-STEP IMPLEMENTATION PROCEDURE FOR LEGACY SYSTEMS

Our software code modification procedure for legacy publication systems proceeds as follows:

- Generate table cells by using the existing procedural language and software
- Locate and discard all marginal table cells, including marginal sensitive cells
- Locate all internal sensitive cells and note their individual protection interval
- At random adjust each sensitive cell value to a lower or upper protection level
- Recreate the marginal cells by re-aggregation of the internal table cells

The recreated table is the public-use table in which all the internal sensitive cell values are kept at a safe distance from the true value.

POTENTIAL APPLICATION BEYOND LEGACY SYSTEMS

A significant amount of research has been done in the recent years to develop safe web-based tabular data release procedures for extremely large, sparse, multi-dimensional tabular data containing sensitive cells. Our proposed method could be used for newly developed multidimensional systems containing both complex hierarchical structure and multiple multidimensional linked structures. For these systems, the dimensionality of each cross section and the total number of cross sections for a public release will need to be decided up front as a part of the system design/users requirement analysis task. After a final decision is made, generating a large number of table cells in multiple multidimensional cross sections and protecting sensitive information in those cross sections using our simplified procedure is a trivial task. In our implementation the innermost marginal cells in all cross sections are treated as internal cells. We later demonstrate the performance of our simplified procedure when applied to seven complex tabular structures available in the public domain.

INDIRECT PROTECTION OF MARGINAL SENSITIVE CELLS

The discussion above addresses protecting internal sensitive cells alone. The sub-additive property of the linear sensitivity measures, used by NSO to arrive at the protection interval, ensures indirect protection of marginal cells, at least in the majority of cases. To accommodate minor exceptions, we propose (1) using randomized variation around the deterministic protection interval and (2) implementing quality control procedures to ensure that the net adjustment for internal sensitive cells in a given dimension exceeds the required adjustment for the marginal sensitive cell. Typically, an adjustment in the direction of change to one internal sensitive cell is needed.

OPTIMAL PLUS/MINUS ALLOCATION

Our simplified procedure targets only non-sensitive marginal cells for adjustments. As a result, the overall percentage of error is kept to a minimum. For the sake of simplicity, we propose that a single dimension consisting of a relatively large number of internal cells (or the dimension of most statistical importance) should be considered for optimal plus/minus allocation of sensitive cells. Such a strategy reduces the possibility of cumulative error adding up to a large value in a dimension containing a large number of internal cells. Based on practical considerations, optimal allocation is not worth the efforts in all other dimensions.

ILLUSTRATIVE EXAMPLE

We use the same 3-D example table used by Dandekar/Cox (2002). The table has 10 columns, 6 rows and 4 levels. The original table cell values are shown in the appendix. The table contains 21 internal and 3 marginal sensitive cells. Using our simplified Legacy CTA procedure, the adjustments to cell values are as follows:

CONTROLLED TABULAR ADJUSTMENTS (10x6x4) TABLE²

--	-35w	--	--	-7w	--	--	--	--	-42
--	--	--	--	--	--	--	--	52w	52
--	--	--	--	--	--	--	--	--	--
--	--	--	--	--	--	--	--	--	--
--	--	--	--	--	--	--	-33w	--	-33
--	-35	--	--	-7	--	--	-33	52	-23
--	53w	--	7w	--	--	--	--	--	60
--	--	--	-78w	--	53w	--	--	--	-25
--	--	--	--	--	-27w	63w	--	--	36
--	--	--	-38w	--	--	--	52w	--	14
--	--	--	--	--	--	36w	--	79w	115
--	53	--	-109w	--	26	99	52	79	200
--	--	--	30w	--	--	54w	-9w	--	75
--	--	--	46w	--	--	--	--	--	46
--	--	--	--	--	--	--	--	-82w	-82
--	32w	--	--	--	--	--	--	--	32
--	--	--	--	--	--	6w	--	--	6
--	32	--	76	--	--	60	-9	-82	77
--	18	--	37	-7	--	54	-9	--	93
--	--	--	-32	--	53	--	--	52	73
--	--	--	--	--	-27	63	--	-82	-46
--	32	--	-38	--	--	--	52	--	46
--	--	--	--	--	--	42	-33w	79w	88
--	50	--	-33	-7	26	159	10	49	254

In the table, the internal sensitive cells are shaded yellow. The three marginal sensitive cells are shaded blue. As mentioned earlier, the protection of marginal sensitive cells follows automatically in the majority of situations. The table showing a percent change in the cell values, after simplified CTA are performed on the marginal cells is included in the appendix. It is apparent from this example that the selection of marginal non-sensitive cells for change offers the maximum benefit in maintaining the overall table quality.

SUMMARY STATISTICS – 7 TEST CASES

We have applied the simplified CTA procedure, without any fine-tuning to protect marginal cells, to seven test cases of varying complexity. These test cases are available in the public domain to all researchers of tabular data protection. The first table below summarizes the characteristics of these tables. In the second table we provide the frequency count of percent cell value change, separately for non-sensitive and sensitive cells, after CTA are implemented. The table also provides the overall average change to non-sensitive cells.

² ‘--’ in a table cell value indicates that no change is made to the cell value

TABULAR DATA TEST CASES IN PUBLIC DOMAIN				
Name	Number of Dimensions	Size	Number of Nonzero Cells	Number of Equality Constraints
HIER13	3-D Hierarchical	13,13,13	2020	3313
HIER16	3-D Hierarchical	16,16,16	3564	5484
BTS4	4-D Hierarchical	54,54,4,4	36570	36310
NINE5D	9-D Linked (Four 5-D Sections)	4,29,3,4,5,6,5,4,5	10733	17295
NINENEW	9-D Linked (Six 4-D Sections)	10,6,6,6,6,6,6,6,6	6546	7340
TWO5IN6	6-D Linked (Two 5-D Sections)	6,4,16,4,4,4	5681	9629
NINE12	9-D Linked (Twelve 4-D Sections)	10,6,6,6,6,6,6,6,6,6	10399	11362

TEST: hier13.cmp

Nonsensitive:1908 Average Abs Dev: 33.785

% from	% To	nonsensitive	Sensitive
.00-	.10	1099	0
.10-	.50	402	0
.50-	1.00	188	0
1.00-	1.50	78	0
1.50-	2.00	45	0
2.00-	5.00	90	61
5.00-	10.00	6	51
10.00-	15.00	0	0
15.00-	30.00	0	0
30.00-	100.00	0	0
100.00-	999.00	0	0

TEST: hier16.cmp

Nonsensitive:3340 Average Abs Dev: 38.442

% from	% To	nonsensitive	Sensitive
.00-	.10	1827	0
.10-	.50	786	0
.50-	1.00	396	0
1.00-	1.50	158	0
1.50-	2.00	68	0
2.00-	5.00	94	83
5.00-	10.00	11	113
10.00-	15.00	0	28
15.00-	30.00	0	0
30.00-	100.00	0	0
100.00-	999.00	0	0

TEST: bts4.cmp

Nonsensitive:34310 Average Abs Dev: 45.36

% from	% To	nonsensitive	Sensitive
.00-	.10	21113	0
.10-	.50	7204	1
.50-	1.00	2835	2
1.00-	1.50	1192	3
1.50-	2.00	704	2
2.00-	5.00	1074	55
5.00-	10.00	185	1480
10.00-	15.00	3	717
15.00-	30.00	0	0
30.00-	100.00	0	0
100.00-	999.00	0	0

TEST: nine5d.cmp

Nonsensitive: 9072 Average Abs Dev: 20.03

% from	% To	nonsensitive	Sensitive
.00-	.10	6580	0
.10-	.50	887	0
.50-	1.00	576	2
1.00-	1.50	348	5
1.50-	2.00	194	1
2.00-	5.00	412	935
5.00-	10.00	75	648
10.00-	15.00	0	70
15.00-	30.00	0	0
30.00-	100.00	0	0
100.00-	999.00	0	0

TEST: ninenew.cmp

Nonsensitive: 5688 Average Abs Dev: 23.91

% from	% To	nonsensitive	Sensitive
.00-	.10	3882	0
.10-	.50	626	1
.50-	1.00	464	2
1.00-	1.50	255	3
1.50-	2.00	141	0
2.00-	5.00	272	512
5.00-	10.00	48	308
10.00-	15.00	0	32
15.00-	30.00	0	0
30.00-	100.00	0	0
100.00-	999.00	0	0

TEST: two5in6.cmp

Nonsensitive: 4961 Average Abs Dev: 25.18

% from	% To	nonsensitive	Sensitive
.00-	.10	3436	0
.10-	.50	539	0
.50-	1.00	363	0
1.00-	1.50	232	3
1.50-	2.00	115	0
2.00-	5.00	233	376
5.00-	10.00	43	320
10.00-	15.00	0	21
15.00-	30.00	0	0
30.00-	100.00	0	0
100.00-	999.00	0	0

TEST: nine12.cmp

Nonsensitive: 9221 Average Abs Dev: 20.80

% from	% To	nonsensitive	Sensitive
.00-	.10	6524	0
.10-	.50	984	1
.50-	1.00	746	1
1.00-	1.50	380	5
1.50-	2.00	209	5
2.00-	5.00	342	686
5.00-	10.00	36	433
10.00-	15.00	0	47
15.00-	30.00	0	0
30.00-	100.00	0	0
100.00-	999.00	0	0

The sensitive cell count within 2 percent of true cell value reflects marginal sensitive cells with either insufficient protection or internal sensitive cells requiring relatively small protection based on the P percent rule. The marginal sensitive cells with insufficient protection could be adequately protected by adjusting the direction of change for one of the internal sensitive cells. This situation typically arises when there are only two internal cells, both of them sensitive and both with a single respondent.

CONCLUSION

The simplified CTA method, unlike the original CTA proposed by Dandekar/Cox (2002), fails to offer multiple options to generate safe tables. However, the simplified procedure, with some additional fine-tuning, offers a cost effective implementation option for legacy systems and newly developed tabular systems without using any special purpose software.

REFERENCES

Dandekar, RA and LH Cox (2002). Synthetic Tabular Data: an alternative to complementary cell suppression for disclosure limitation of tabular data. Manuscript.

APPENDIX

ORIGINAL (10, 6, 4) TABLE FROM DANDEKAR/COX(2002)

6764	714w	3356	4067	140w	--	3932	1478	--	20451
1994	--	5593	--	3022	3504	--	3220	1042w	18375
3744	--	3708	--	3678	2502	--	--	--	13632
2810	10632	--	2445	--	--	2313	2978	7548	28726
3682	--	--	--	4667	1988	1748	664w	--	12749

18994	11346	12657	6512	11507	7994	7993	8340	8590	93933
--	539w	--	70w	--	7472	715	3832	--	12628
2253	--	4948	786w	472	1074w	1830	5030	--	16393
640	--	986	--	--	544w	631w	48	750	3599
1334	--	1016	382w	3175	3302	3803	1050w	--	14062
1648	2814	--	--	--	2102	726w	--	1598w	8888

5875	3353	6950	1238w	3647	14494	7705	9960	2348	55570
--	3552	3476	614w	1916	1131	549w	92w	1772	13102
--	--	3222	928w	--	--	308	429	87	4974
4145	--	--	3692	2115	4196	414	3804	820w	19186
5995	644w	--	--	2410	1677	--	1912	4134	16772
2016	--	--	2212	2826	1627	134w	--	--	8815

12156	4196	6698	7446	9267	8631	1405	6237	6813	62849

6764	4805	6832	4751	2056	8603	5196	5402	1772	46181
4247	--	13763	1714	3494	4578	2138	8679	1129	39742
8529	--	4694	3692	5793	7242	1045	3852	1570	36417
10139	11276	1016	2827	5585	4979	6116	5940	11682	59560
7346	2814	--	2212	7493	5717	2608	664w	1598w	30452

37025	18895	26305	15196	24421	31119	17103	24537	17751	212352

PERCENT CONTROLLED TABULAR ADJUSTMENTS

--	-4.9%	--	--	-5.0%	--	--	--	--	-0.2%
--	--	--	--	--	--	--	5.0%	--	.3%
--	--	--	--	--	--	--	--	--	--
--	--	--	--	--	--	--	--	--	--
--	--	--	--	--	--	--	-5.0%	--	-0.3%
=====									
--	-0.3%	--	--	-0.1%	--	--	-0.4%	.6%	.0%
--	9.8%	--	10.0%	--	--	--	--	--	.5%
--	--	--	-9.9%	--	4.9%	--	--	--	-0.2%
--	--	--	--	--	-5.0%	10.0%	--	--	1.0%
--	--	--	-9.9%	--	--	--	5.0%	--	.1%
--	--	--	--	--	--	5.0%	--	4.9%	1.3%
=====									
--	1.6%	--	-8.8%	--	.2%	1.3%	.5%	3.4%	.4%
--	--	--	4.9%	--	--	9.8%	-9.8%	--	.6%
--	--	--	5.0%	--	--	--	--	--	.9%
--	--	--	--	--	--	--	--	-10.0%	-0.4%
--	5.0%	--	--	--	--	--	--	--	.2%
--	--	--	--	--	--	4.5%	--	--	.1%
=====									
--	.8%	--	1.0%	--	--	4.3%	-0.1%	-1.2%	.1%
--	.4%	--	.8%	-0.3%	--	1.0%	-0.2%	--	.2%
--	--	--	-1.9%	--	1.2%	--	--	4.6%	.2%
--	--	--	--	--	-0.4%	6.0%	--	-5.2%	-0.1%
--	.3%	--	-1.3%	--	--	--	.9%	--	.1%
--	--	--	--	--	--	1.6%	-5.0%	4.9%	.3%
=====									
--	.3%	--	-0.2%	.0%	.1%	.9%	.0%	.3%	.1%

