Comparative Evaluation of Four Different Sensitive Tabular Data Protection Methods Using a Real Life Table Structure of Complex Hierarchies and Links Populated with Artificial Data

Ramesh A. Dandekar
Energy Information Administration
Washington DC

(Ramesh.Dandekar@EIA.DOE.GOV)

UNECE2007 – 17-19 December 2007

Tabular Data Protection Methods*

- Classical LP Based Cell Suppression
- Network Flow Based Cell Suppression (USBC)
- LP Based Synthetic Tabular Data / CTA (Dandekar 2001)
- Micro Data Level Noise Addition (USBC)

P = 10 % Rule Used

* Uses Proprietary Research Tools

Two Three Dimensional *HYPOTHETICAL* Tables Linked in Four Dimensional Space

1st Table: "Volumes *by Grade*, Sales Type, PAD District, and State" ,

and

2nd Table: "Volumes *by Formulation*, Sales Type, PAD District, and State"

First Table

TABLE 1: VOLUMES BY *GRADE*, SALES TYPE, PAD DISTRICT, AND STATE

Geographic Area	Regular				Mid-grade				Premium				All Grades			
	DTW	R A C K	B U L K	T O T A L	DTW	R A C K	B U L K	T O T A L	DTW	R A C K	B U L K	T O T A L	DTW	R A C K	B U L K	T O T A L
United States	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx

Second Table

TABLE 2: VOLUMES BY *FORMULATION*, SALES TYPE, PAD DISTRICT, AND STATE

Geographic Area	Conventional				Oxygenated				Reformulated				All Formulations			
	DTW	R A C K	B U L K	T O T A L	DTW	R A C K	B U L K	T O T A L	DTW	R A C K	B U L K	T O T A L	DTW	R A C K	B U L K	T O T A L
United States	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx

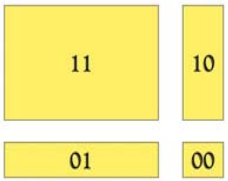
1st Table Grades

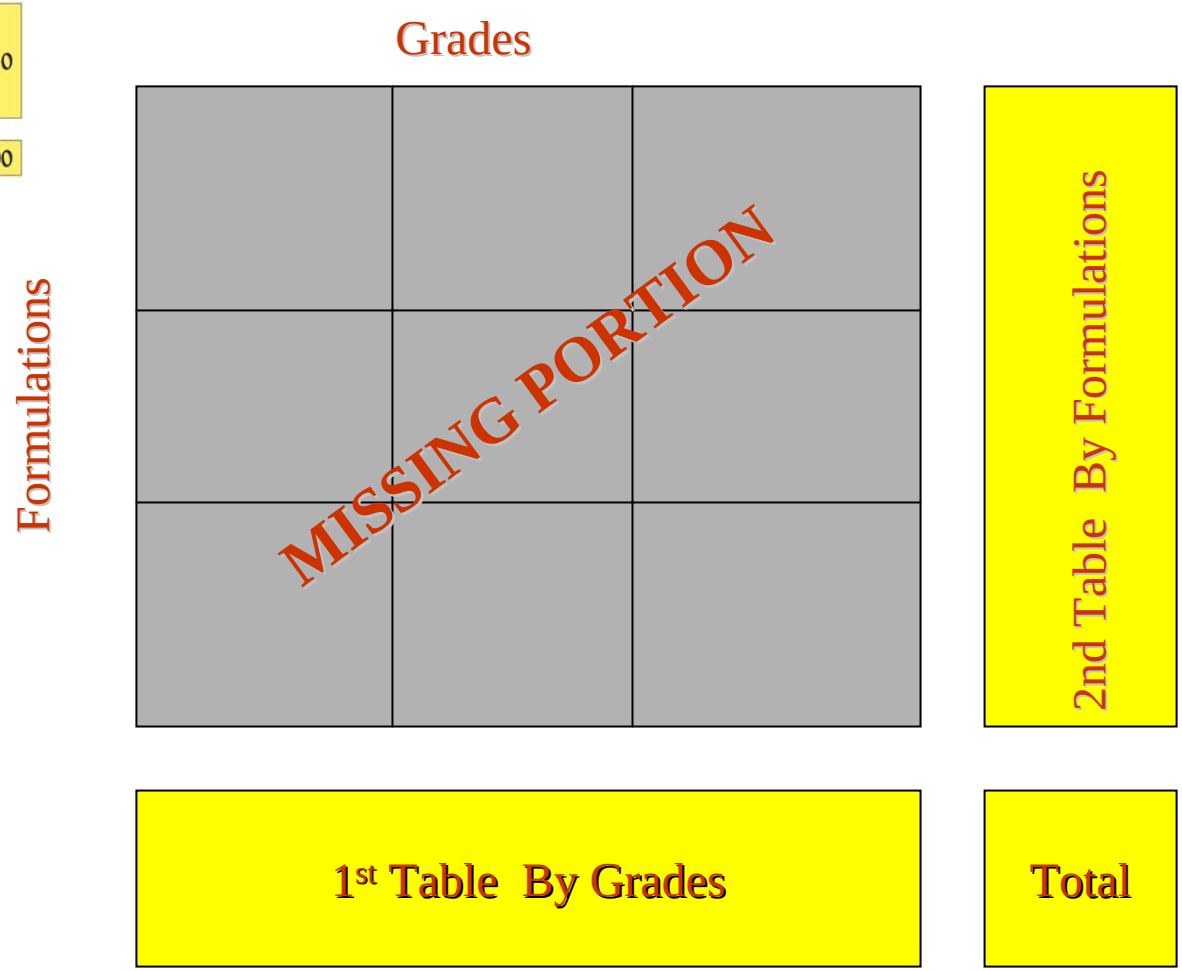
- Regular
- Midgrade
- Premium
- Total All Grades

2nd Table Formulations

- Conventional
- Oxygenated
- Reformulated
- Total All Formulations

Guidance Matrix Components for 2-D Table





Four Layers: 1) DTW 2) Rack 3) Bulk 4) Total

Corresponding to each PAD, State and US Level Cell

1,000 Synthetic Micro Data Records Containing Six Variables

Four Categorical Variables

- 51 States
- 3 Grade Types
- 3 Sale Types
- 3 Formulation Types

One Magnitude Variable

One Sample Weight Variable

Classical LP-Based Cell Suppresion vs CTA

Classical LP/CTA 01 regular

	DTW	Rack	Bulk	Total	←--- CTA Solution --→			
United States	188668.0	218471.0	170021.0	577160.0	-130.A	113.A	-61.A	-78.A
PAD District I	64625.0	72994.0	65620.0	203239.0	-8.A	69.A	-143.A	-82.A
Subdistrict IA	25314.0	28780.0	16952.0	71046.0	-8.A	8.A	0.	0.
Connecticut	6258.0	1494.0	1700.0	9452.0	0.	0.	0.	0.
Maine	3936.0	4719.0	4429.0	13084.0	0.	0.	0.	0.
Massachusetts	172.0 w	3840.0 s	.0	4012.0	-8.w	8.A	0.	0.
New Hampshire	7879.0	.0	3188.0	11067.0	0.	0.	0.	0.
Rhode Island	1748.0 s	6224.0 s	3976.0	11948.0	0.	0.	0.	0.
Vermont	5321.0	12503.0	3659.0	21483.0	0.	0.	0.	0.
Subdistrict IB	19417.0	16493.0	22335.0	58245.0	0.	48.A	-61.A	-13.A
Delaware	6978.0	2400.0	4272.0	13650.0	0.	48.A	0.	48.A
District of Columbia	2253.0	5070.0	11338.0	18661.0	0.	0.	0.	0.
Maryland	3311.0	1836.0 s	1079.0 w	6226.0	0.	0.	-60.w	-60.A
New Jersey	6875.0	.0	144.0	7019.0	0.	0.	0.	0.
New York	.0	648.0 s	784.0 w	1432.0	0.	0.	-39.w	-39.A
Pennsylvania	.0	6539.0	4718.0	11257.0	0.	0.	38.A	38.A
Subdistrict IC	19894.0	27721.0	26333.0	73948.0	0.	13.A	-82.A	-69.A
Florida	.0	10857.0	1847.0	12704.0	0.	0.	-17.A	-17.A
Georgia	9961.0	.0	.0	9961.0	0.	0.	0.	0.
North Carolina	2268.0 s	7226.0 s	8464.0 s	17958.0	0.	13.A	-65.A	-52.A
South Carolina	1195.0	5887.0	7582.0	14664.0	0.	0.	0.	0.
Virginia	3560.0	.0	3625.0	7185.0	0.	0.	0.	0.
West Virginia	2910.0 s	3751.0 s	4815.0 s	11476.0	0.	0.	0.	0.
PAD District II	76174.0	62147.0	54796.0	193117.0	-71.A	0.	126.A	55.A
Illinois	4128.0	.0	.0	4128.0	0.	0.	0.	0.
Indiana	4613.0 s	.0	3846.0 s	8459.0 s	-14.A	0.	14.A	0.
Iowa	1149.0	4196.0	4216.0 s	9561.0 s	0.	0.	0.	0.
Kansas	11996.0	10330.0	1948.0	24274.0	-57.A	0.	112.A	55.A
Kentucky	5826.0 s	2787.0 s	6523.0	15136.0	0.	0.	0.	0.
Michigan	2022.0 s	.0	6668.0 s	8690.0	0.	0.	0.	0.
Minnesota	6400.0	3694.0	1332.0	11426.0	0.	0.	0.	0.
Missouri	5915.0	10385.0	3934.0	20234.0	0.	0.	0.	0.
Nebraska	2652.0	7667.0	942.0	11261.0	0.	0.	0.	0.
North Dakota	4671.0	8286.0	.0	12957.0	0.	0.	0.	0.
Ohio	7197.0	.0	3477.0	10674.0	0.	0.	0.	0.
Oklahoma	4030.0	1864.0	4339.0	10233.0	0.	0.	0.	0.
South Dakota	24.0	11013.0	5526.0	16563.0	0.	0.	0.	0.
Tennessee	2242.0	645.0	8325.0	11212.0	0.	0.	0.	0.
Wisconsin	13309.0	1280.0 s	3720.0 s	18309.0	0.	0.	0.	0.
PAD District III	15248.0	23726.0	26417.0	65391.0	0.	-19.A	0.	-19.A
Alabama	3504.0	259.0 w	2856.0 s	6619.0	0.	25.w	0.	25.A
Arkansas	1598.0	5628.0	6358.0	13584.0	0.	0.	0.	0.
Louisiana	.0	3088.0 s	4667.0 s	7755.0	0.	0.	0.	0.
Mississippi	666.0	8925.0	2980.0	12571.0	0.	0.	0.	0.
New Mexico	8410.0	4928.0	6696.0	20034.0	0.	0.	0.	0.
Texas	1070.0	898.0 w	2860.0 s	4828.0	0.	-44.w	0.	-44.A
PAD District IV	13561.0	23112.0	8479.0	45152.0	-51.A	132.A	-44.A	37.A
Colorado	.0	8772.0	5637.0	14409.0	0.	0.	0.	0.
Idaho	925.0 s	940.0 w	890.0 w	2755.0 s	0.	94.w	-44.w	50.A
Montana	514.0 w	7358.0 s	.0	7872.0 s	-51.w	0.	0.	-51.A
Utah	5676.0 s	382.0 w	.0	6058.0 s	0.	38.w	0.	38.A
Wyoming	6446.0 s	5660.0 s	1952.0 s	14058.0	0.	0.	0.	0.
PAD District V	19060.0	36492.0	14709.0	70261.0	0.	-69.A	0.	-69.A
Alaska	.0	7948.0	4300.0	12248.0	0.	0.	0.	0.
Arizona	2721.0	828.0	2189.0	5738.0	0.	0.	0.	0.
California	3792.0	3728.0	2251.0	9771.0	0.	-69.A	0.	-69.A
Hawaii	1038.0	6141.0	327.0	7506.0	0.	0.	0.	0.
Nevada	2555.0	3522.0	.0	6077.0	0.	0.	0.	0.
Oregon	.0	14325.0	3040.0	17365.0	0.	0.	0.	0.
Washington	8954.0	.0	2602.0	11556.0	0.	0.	0.	0.

DiAna Cell Suppression Pattern vs Noise Addition

Diana/Noise 01 regular

	DTW		Rack		Bulk		Total		<--- Census' Noise Method-->			
United States	188668.		218471.		170021.		577160.		-16.	357.	81.	422.
PAD District I	64625.		72994.		65620.		203239.		-144.	219.	-53.	22.
Subdistrict IA	25314.		28780.		16952.		71046.		-267.	-115.	83.	-300.
Connecticut	6258.	C	1494.	C	1700.		9452.		-161.	-26.	-35.	-222.
Maine	3936.		4719.		4429.		13084.		150.	-69.	64.	145.
Massachusetts	172.	P	3840.	C	.		4012.		-9.	-65.	0.	-74.
New Hampshire	7879.		.		3188.		11067.		-159.	0.	81.	-78.
Rhode Island	1748.		6224.		3976.		11948.		-53.	143.	79.	168.
Vermont	5321.	C	12503.	C	3659.		21483.		-34.	-98.	-106.	-238.
Subdistrict IB	19417.		16493.		22335.		58245.		123.	51.	-137.	38.
Delaware	6978.	C	2400.	C	4272.		13650.		133.	19.	47.	200.
District of Columbia	2253.		5070.		11338.		18661.		-53.	85.	86.	118.
Maryland	3311.	C	1836.	C	1079.	P	6226.		-120.	-46.	-110.	-275.
New Jersey	6875.	C	.		144.	C	7019.		163.	0.	5.	168.
New York	.		648.	C	784.	P	1432.		0.	22.	-41.	-19.
Pennsylvania	.		6539.		4718.		11257.		0.	-30.	-124.	-154.
Subdistrict IC	19894.		27721.		26333.		73948.		0.	283.	1.	284.
Florida	.		10857.	C	1847.	C	12704.		0.	15.	-15.	1.
Georgia	9961.		.		.		9961.		93.	0.	0.	93.
North Carolina	2268.		7226.	C	8464.	C	17958.		94.	205.	-49.	250.
South Carolina	1195.		5887.		7582.		14664.		25.	-5.	165.	186.
Virginia	3560.		.		3625.		7185.		-187.	0.	-112.	-298.
West Virginia	2910.		3751.	C	4815.	C	11476.		-26.	68.	10.	52.
PAD District II	76174.		62147.		54796.		193117.		12.	186.	22.	220.
Illinois	4128.		.		.		4128.		154.	0.	0.	154.
Indiana	4613.		.		3846.	C	8459.	C	158.	0.	-112.	46.
Iowa	1149.	C	4196.	C	4216.	C	9561.		50.	-49.	60.	62.
Kansas	11996.		10330.		1948.		24274.		-62.	-156.	21.	-197.
Kentucky	5826.	C	2787.	C	6523.	C	15136.		106.	61.	-73.	94.
Michigan	2022.	C	.		6668.	C	8690.	C	73.	0.	-158.	-85.
Minnesota	6400.		3694.		1332.		11426.		-145.	110.	37.	2.
Missouri	5915.		10385.		3934.		20234.		-195.	46.	13.	-136.
Nebraska	2652.		7667.		942.		11261.		-80.	148.	32.	100.
North Dakota	4671.		8286.		.		12957.		30.	-18.	0.	12.
Ohio	7197.		.		3477.		10674.		-101.	0.	-40.	-141.
Oklahoma	4030.		1864.		4339.		10233.		46.	-52.	82.	76.
South Dakota	24.		11013.		5526.		16563.		1.	75.	94.	170.
Tennessee	2242.	C	645.	C	8325.		11212.		-70.	-14.	157.	73.
Wisconsin	13309.		1280.		3720.		18309.		47.	35.	-91.	-9.
PAD District III	15248.		23726.		26417.		65391.		30.	-233.	181.	-22.
Alabama	3504.	C	259.	P	2856.		6619.		-94.	-26.	-51.	-171.
Arkansas	1598.		5628.		6358.		13584.		11.	-43.	80.	48.
Louisiana	.		3088.	C	4667.	C	7755.		0.	-21.	115.	94.
Mississippi	666.	C	8925.	C	2980.		12571.		-23.	-218.	-155.	-396.
New Mexico	8410.		4928.		6696.		20034.		158.	121.	111.	389.
Texas	1070.	C	898.	P	2860.	C	4828.		-22.	-46.	81.	14.
PAD District IV	13561.		23112.		8479.		45152.		-106.	-133.	228.	-11.
Colorado	.		8772.		5637.		14409.		0.	-116.	129.	13.
Idaho	925.	C	940.	P	890.	P	2755.	C	-19.	100.	49.	131.
Montana	514.	P	7358.	C	.		7872.		53.	-23.	0.	30.
Utah	5676.	C	382.	P	.		6058.		-99.	-39.	0.	-138.
Wyoming	6446.		5660.	C	1952.	C	14058.	C	-41.	-54.	49.	-46.
PAD District V	19060.		36492.		14709.		70261.		192.	318.	-297.	213.
Alaska	.		7948.		4300.		12248.		0.	36.	-87.	-51.
Arizona	2721.	C	828.	C	2189.	C	5738.		82.	29.	-67.	44.
California	3792.		3728.	C	2251.	C	9771.		-84.	97.	-61.	-48.
Hawaii	1038.	C	6141.		327.	C	7506.		76.	-196.	11.	-108.
Nevada	2555.	C	3522.	C	.		6077.		-19.	56.	0.	37.
Oregon	.		14325.		3040.		17365.		0.	297.	-64.	233.
Washington	8954.		.		2602.		11556.		136.	0.	-30.	106.

Comparative Evaluation of Cell Suppression Methods

Classical Cell Suppression

- 294 Suppressions
- Sensitive Cells Fully Protected

Network Flow Method

- 479 Suppressions
- Sensitive Cells Fully Protected
- 3 exact Disclosures of non-sensitive cells

CTA vs NOISE - TABULAR DATA QUALITY

CTA frequency Distribution

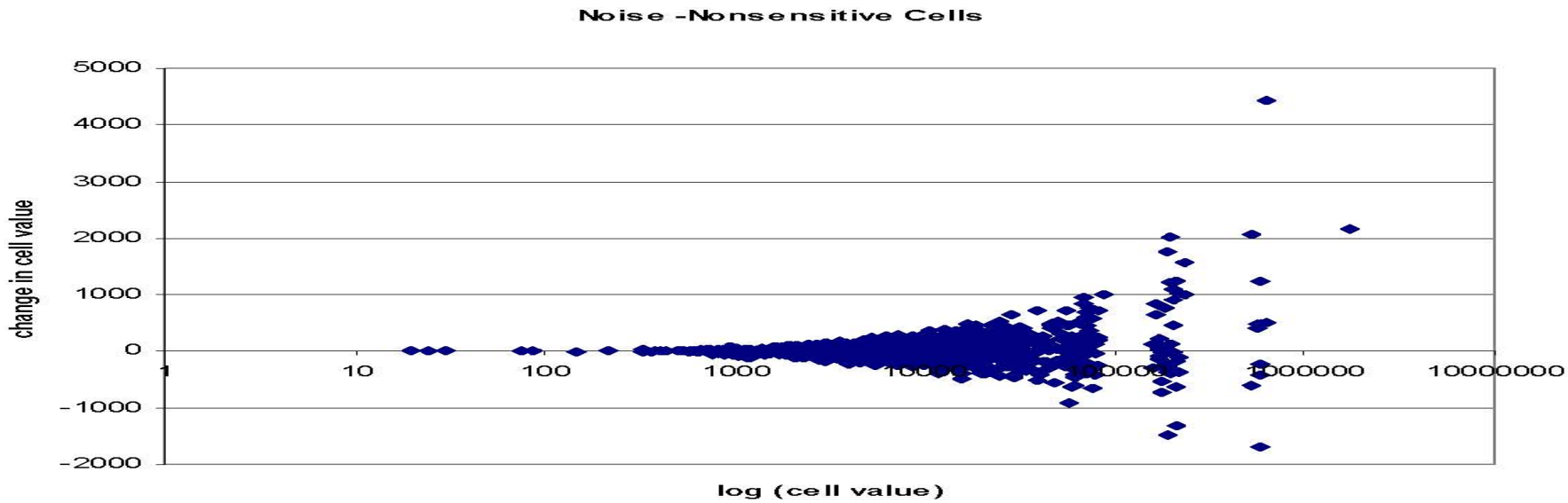
% From	% To	Non-Sensitive	Sensitive
.00 -	.10	1235	0
.10 -	.50	137	1
.50 -	1.00	60	0
1.00 -	1.50	15	0
1.50 -	2.00	13	1
2.00 -	5.00	15	50
5.00 -	10.00	3	26
10.00 -	15.00	0	0
15.00 -	30.00	0	0
30.00 -	100.00	0	0

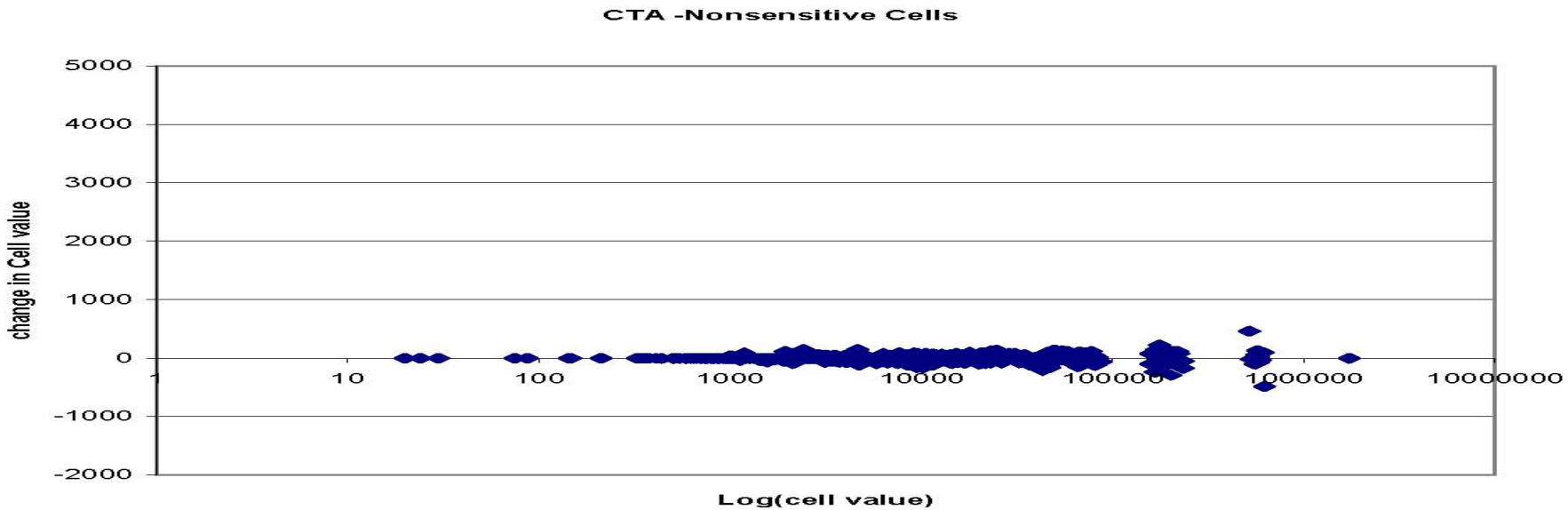
1432

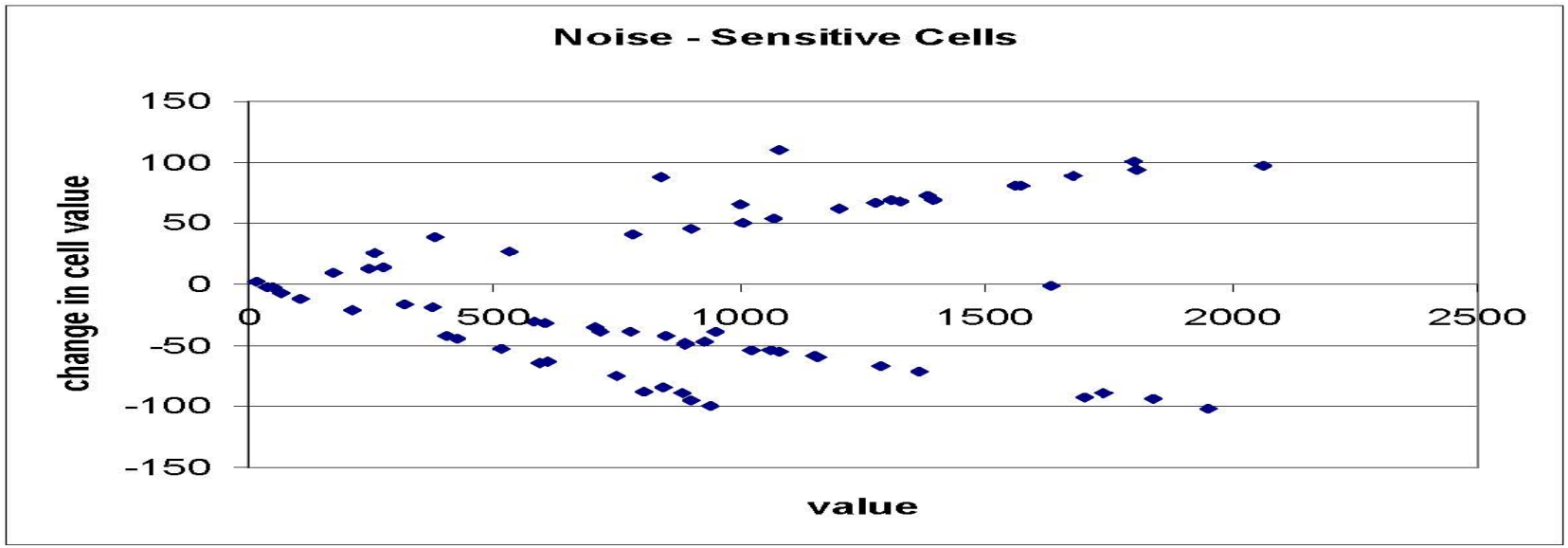
Noise Frequency Distribution

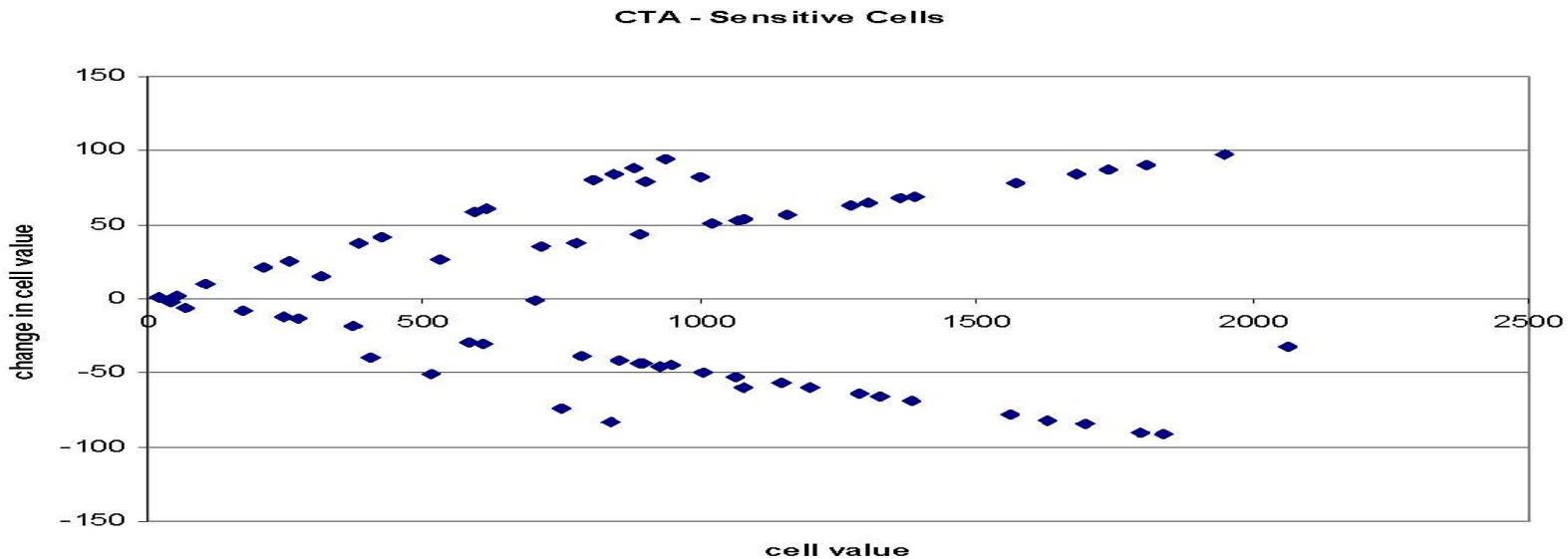
% From	% To	Nonsensitive	Sensitive
.00-	.10	96	1
.10-	.50	272	0
.50-	1.00	265	0
1.00-	1.50	215	0
1.50-	2.00	164	0
2.00-	5.00	439	2
5.00-	10.00	27	51
10.00-	15.00	0	24
15.00-	30.00	0	0
30.00-	100.00	0	0

633


Noise -Nonsensitive Cells
5000
4000
3000
2000
1000
0
-1000
-2000
change in cell value
1
10
100
1000
10000
100000
1000000
10000000
log (cell value)


CTA -Nonsensitive Cells
5000
4000
3000
2000
1000
0
-1000
-2000
change in Cell value
1
10
100
1000
10000
100000
1000000
10000000
Log(cell value)


Noise - Sensitive Cells
change in cell value
150
100
50
0
-50
-100
-150
0
500
1000
1500
2000
2500
value


CTA - Sensitive Cells
change in cell value
150
100
50
0
-50
-100
-150
0
500
1000
1500
2000
2500
cell value

THANK YOU!

ADDITIONAL INFORMATION FROM

http://mysite.verizon.net/vze7w8vk/