[IN]APPROPRIATE USE OF STATISTICAL MEASURES IN [THE NAME OF] BALANCING DATA QUALITY AND CONFIDENTIALITY OF TABULAR FORMAT MAGNITUDE DATA[1]

Ramesh A. Dandekar

Energy Information Administration, U. S. Department of Energy, Washington DC 20585
Ramesh.dandekar@eia.gov { http://mysite.verizon.net/vze7w8vk/ }

Abstract. Statisticians are aware of the fact that measures such as: mean, variance, Pearson correlation coefficient are disproportionately influenced by relatively few extremely large observations and, therefore, are unreliable as statistical measures in comparing overall quality of data with an extremely skewed distribution. Tabular data cells follow an extremely skewed distribution. In this paper we show that linear-programming-based controlled tabular adjustments (CTA), which generates synthetic tabular data (Dandekar2001), makes use of a least absolute difference linear regression model and is well-suited to control overall data quality on its own without additional steps proposed by quality preserving controlled tabular adjustments (QP-CTA) that has been heavily promoted to the statistical community since 2003.

Key words: L1 norm regression, disclosure control, synthetic tabular data

1 Introduction

Controlled tabular adjustments to generate synthetic tabular data (Dandekar2001) is the only method which has been currently demonstrated to be effective in protecting sensitive cells in tables containing extremely large multi-dimensional counts data and magnitude data with complex hierarchies and linked table structures. The *concept* of synthetic tabular data, which closely mimics real tabular data and at the same time protects sensitive information contained in the tabular format magnitude data, was introduced in December of 1996 as an alternative method to the then state-of-the-art micro data level multiplicative noise addition method. The synthetic tabular data generation procedure has been perceived to be a non-statistical ad hoc procedure by many in the statistical community. In this paper we demonstrate that the CTA procedure to generate synthetic tabular data, as formally proposed (Dandekar2001), is in fact a valid statistical procedure and utilizes a linear regression error correction model to achieve that goal. We also demonstrate that the original procedure is well-suited to control overall data quality on its own without incorporating the additional steps proposed by quality preserving controlled tabular adjustments (QP-CTA), since 2003. To demonstrate the statistical properties of the CTA protected synthetic tabular data, we use a real life table structure of typical hierarchical and linked complexities (Dandekar2007) populated with an artificial micro data generator developed and used by this author since 1998.

[1] 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).

2 Fundamentals of Tabular Format Magnitude Data

Tabular format magnitude data is not micro data. Each table cell within a table is *macro data* and consists of an aggregate of multiple micro data records. A table, in general, is a collection of individual table cells and often contains many *multiple levels of* aggregates of *"aggregates of macro data"* resulting from imbedded hierarchical table structures. The concept of macro data analysis, therefore, is not identical to the concept of micro data analysis. Statistical measures, such as average value, standard deviation, and correlation coefficient analysis of micro data (records), have different interpretations in macro data analysis. Multiple levels of aggregates of "*aggregates of macro data*" further dampen the properties associated with individual micro data records. ***Exceptions to this rule are individual table cells with very few micro data records, and table cells dominated by a few micro data records. This phenomenon necessitates the creation of synthetic tabular data, in the first place.***

Tabular data cells within a table follow a skewed distribution. This property is further aggravated significantly with the increase in the dimensionality and the complexity of table structure and by groups of logically related (linked) table structures that need to be analyzed and protected together. As a result, measures such a mean, standard deviation, and correlation coefficients that are commonly used for symmetrically distributed data are less useful in comparing synthetic tabular data with real tabular data. The more appropriate comparison measures are by comparing either probability density functions or cumulative distribution functions of synthetic tabular cells with real table cells.

3 Connection of Mean and Median to Optimization

The **s**tatistical mean value of a variable is widely defined as the sum of multiple observations divided by the total number of observations. Similarly, median value is widely defined as the value of a variable at the mid-point when all observations are arranged in non-decreasing sequence. Strictly speaking though, the correct technical definition of mean value is the point that minimizes the sum of the squared deviations between the data points and mean value, itself[2]. In short, the mean is estimated by using ***a least squares measure***. Similarly, the correct technical definition of median value is the point that minimizes the sum of the absolute values of the difference between each data point and the median, itself. Thus, the median is estimated by using ***a least absolute deviation measure***.

For a symmetric distribution, the mean and median are the same. As a result, the outcome from the least squares measure is similar to the least absolute deviation measure. Similarly, for a skewed distribution, the median is a more accurate representation of central tendency than the mean. Consequently, the outcome from the least absolute deviation measure, which is used for generating synthetic tabular data (Dandekar2001), is more appropriate and robust measure than the outcome from the least square measures proposed by QP-CTA.

[2] Vanderbei, Robert, "Linear Programming Foundations and Extensions", *Springer International Series in Operations Research & Management Science, Vol. 114.*

4 Typical Linear Regression Model

Typical linear model is of the form

$$e = [\, Y - \sum^{n} a_j * X_j \,] \qquad (1)$$

Where Y = response variable, a_j = unknown parameter vector, X_j = controlled variable vector, n = number of control variables and e = estimation error.

By using m different observations available for Y and X_j, the objective is to determine the best values of parameter vector a_j which will estimate Y with overall minimum estimation error. The overall estimation error e in the equation (1) can be minimized by using a linear regression model either by using the L1 norm measure, or by using the L2 norm measure.

By using the L1 norm measure the linear regression model in the error form (1) becomes:

$$\text{argmin} \left(\text{ABS} [\, \sum^{m} e_i \,] \right) = \text{argmin} \sum^{m} \left(\text{ABS} [\, Y_i - \sum^{n} a_j * X_{ij} \,] \right) \qquad (2)$$

Similarly, by using the L2 norm measure the linear regression model in the error form (1) becomes:

$$\text{argmin} \sum^{m} (e_i)^2 = \text{argmin} \sum^{m} [\, Y_i - \sum^{n} a_j * X_{ij} \,]^2 \qquad (3)$$

5 Comparison of the CTA Model to Linear Regression Model

The synthetic tabular data generation procedure is formulated in the Dandekar2001 paper by using Fischetti and Salazar notations and mixed integer linear programming (MILP) formulation as:

$$\text{Minimize} \sum [\, c_i (y_i^{+} + y_i^{-}) \,] \qquad (4)$$

Subject to

$$M (y_i^{+} - y_i^{-}) = 0 \qquad (5)$$

$$0 =< y_i^{+} =< UB_i \qquad (6)$$

$$0 =< y_i^{-} =< LB_i \qquad (7)$$

$$y_{ik}^{+} >= BOUND_{ik} * I_{ik} \qquad (8)$$

$$y_{ik}^{-} >= BOUND_{ik} * (1 - I_{ik}) \qquad (9)$$

where

I_{ik} is a binary zero/one variable

$BOUND_{ik}$ is confidentiality bound for sensitive cell ik

ik (k = 1, ... p) p sensitive cells

i = 1, ... n n non-zero table cells

y_i^{+} = positive adjustment to cell value

y_i^- = negative adjustment to cell value

UB_i and LB_i Upper and lower cell bounds

c_i = cost function.

The conditions (5) to (9) are imposed in an attempt to maintain table cell additivity by using (5); to maintain tabular data quality by using (6) and (7); and to prevent statistical disclosure of sensitive tabular cell values by imposing (8) and (9).

The proposed CTA model used to generate synthetic tabular data, is a least absolute deviation linear regression model[3] ***similar*** *to* shown in equation (2) with one difference[4]. In the CTA formulation ***the error correction terms are associated with the "n" control variables and not with the "m" observations***[5]. The unknown regression parameter a_j in the CTA model in essence also plays a role of error correction term. As a result, equation (2) is processed separately in the CTA formulation by decomposing it in to two different components, namely an error minimization component (4) and table additivity component (5). Each controlled tabular adjustment is treated as a separate regression parameter a_j. The unknown parameters are estimated by performing a linear regression analysis. The regression parameter a_j is further separated in to two independent components, y_j^+ and y_j^-, by using a mathematical relationship of the form $a_j = y_j^+ - y_j^-$ to separately identify positive adjustment and negative adjustment to the table cell value. In this formulation y_j^+ and y_j^- are greater than or equal to zero. As a result, when $y_j^+ > 0$, then $y_j^- = 0$; and when $y_j^- > 0$, then $y_j^+ = 0$.

The control variable X_{ij} from (2) is used to represent the table structure in a matrix notations M in (5) by using a discrete variable {+1, 0, -1} to define m different additive table relationships. The control variable is assigned a value of zero when the table cell is not contributing to any given additive table relationship. The control variable is assigned a value of +1 when the table cells makes an additive contribution to a given table relationship. The value of -1 is assigned to the control variable when the table cell value is an aggregate of other cell values contributing to the additive relationship. The response variable Y_i is always equal to zero to ensure that table additivity conditions are fully satisfied. Imbalances imposed in the table structure (5) by non-zero (conditional) table parameters (8) and (9) initiates the error minimization task to rebalance the entire table.

In the final LP solution, not all controlled variables enter as explanatory variables. Similarly, in the final solution, not all the equality constraints (additive relations) come into play. The number of controlled variables selected as explanatory variables (regression parameter not equal to zero)[6] are of the same order of magnitude as the number of binding additive table relations.

---

[3] For additional technical details on the use of linear programming to perform linear regression analysis please refer to a technical paper by Harvey M. Wagner, " Linear Programming Techniques for Regression Analysis", Journal of the American Statistical Association, Vol. 54, No. 285 (Mar., 1959), pp. 206-212.

[4] For additional explanation on CTA formulation with Wagner1959 paper please see appendix A.

[5] CTA can also be formulated as shown in equation (2) by associating error correction terms with "m" observations. That formulation, however, is less flexible.

[6] In the linear programming based tabular data complementary cell suppression (CCS) procedure, table cells with non-zero regression parameters are suppressed. Unlike CTA procedure, the CCS procedure uses multiple localized linear regression model runs to develop cell suppression pattern. **In short CCS procedure is a statistical procedure**.

6 Modified Interpretation of Cost Functions in the CTA Model

The "appropriate statistical" measure of the overall quality of synthetic tabular data generated by the CTA model could be derived by aggregation of total percent error introduced in all table cell values. This objective is best achieved by measuring the cumulative absolute percent (or fractional) error introduced by the synthetic data generation procedure over all table cell values. That requirement is easily implemented in equation (4) by an appropriate use of the cost function c_i in the linear programming objective function. Directly associating the cost function c_i, which is a reciprocal of table cell value x_i, with deviations in table cell value variables (y_i^+ and y_i^-) appearing in the objective function (4), in essence converts the deviations into percent (or fractional) error form. The variable transformation of $y_{new}^+ = y^+ / x$ and $y_{new}^- = y^- / x$ allows converting the weighted CTA model specified in equation (4) into an un-weighted least absolute deviation linear regression model.

In the LP-based tabular data complementary cell suppression procedure related literature, the cost functions are typically associated with deviation in the table cell values. Cost functions such as, 1) constant value, 2) cell value, 3) log (cell value), and 4) reciprocal of cell value are often used to develop different cell suppression patterns. This interpretation of the cost function needs to be changed when the same cost functions are used to generate synthetic tabular data by controlled tabular adjustments. The change in the interpretation of the cost function could be achieved by transformation of the cost function to a new cost function by simultaneously multiplying and dividing the cost function by a related table cell value, and then by associating the cell value in the denominator of the cost function with the variable in the objective function to transform the variable into fractional (or percent) form. Such a transformation allows us to estimate the cumulative percent error associated with the CTA procedure. Table 1 show the appropriate interpretations of some of the conventional LP cost functions, when they are used to generate synthetic tabular data by controlled tabular adjustments

Table 1.

CONVENTIONAL LP COST FUNCTION	EQUIVALENT CTA RELATED PERCENT CHANGE COST FUNCTION
CONSTANT	CELL VALUE
CELL VALUE	CELL VALUE * CELL VALUE = (CELL VALUE)^2
LOG (CELL VALUE)	LOG (CELL VALUE) * CELL VALUE
1 / CELL VALUE	CONSTANT
1 / (CELL VALUE)^2	1 / CELL VALUE
[LOG(CELL VALUE)] / CELL VALUE	LOG(CELL VALUE)

7 Statistical Properties of the CTA Model

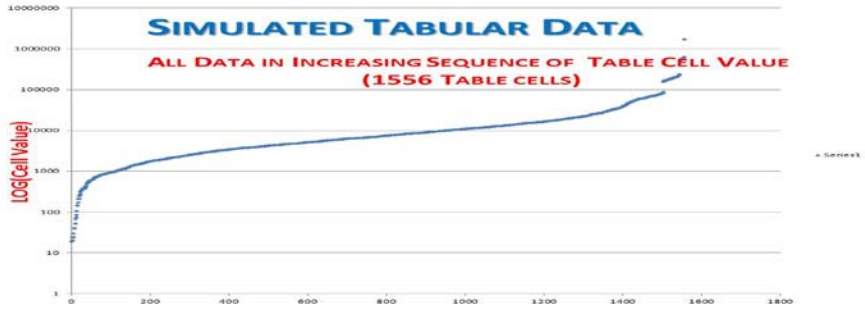
Synthetic tabular data is created to protect sensitive table cells. Sensitive cells typically have small magnitudes and are clustered towards the left of the probability density function. The intentional distortion in the sensitive table cell values is counter balanced by a combination of the direction of distortion (up or down) of other sensitive cell values and by a "minor" adjustment of non-sensitive cell values. Various options for controlled

tabular adjustments by an appropriate use of the cost function in (4) are possible. Multiple factors such as frequency of table publication, and table structure determine which option to use. Irrespective of these factors, the major requirement for the synthetic tabular data generation procedure is that the least amount of table cells undergo changes in their value, no matter how trivial the change. This requirement is easy to accomplish when the CTA model estimation errors follow a Laplace distribution with a relatively large peak and rapid decay.

The CTA model uses a linear error model over a relatively narrow range of the explanatory variables and therefore avoids problems associated with extremely wide and skewed table cell distribution. The typical CTA model estimation range is from zero percent change (no change) to up to 100% change in the table cell value. To demonstrate the effectiveness of the CTA model, we have used the same real life table structure of moderate hierarchical and linked complexity used in Dandekar2007[7]. The table consists of eight two-dimensional cross sections (two three-dimensional cross sections) linked in the four-dimensional space. The table is populated with the non-real synthetic micro data using the same procedure described in that paper. The p percent rule with p=10% is used to identify sensitive cells.

The distribution associated with simulated tabular data and real tabular data (obtained by using a published data table in February 2006) are shown in Fig 1 and Fig 2. The simulated tabular data consists of 1,556 table cells while the real table contains 1,125 table cells. In both graphs the vertical scale uses a log transform of the table cell value. We have shown the distribution from real table cell values along side with simulated data used in this paper to demonstrate the effectiveness of the artificial synthetic micro data generator used by this author since 1998 to recreate real life tables for SDL research projects[8].

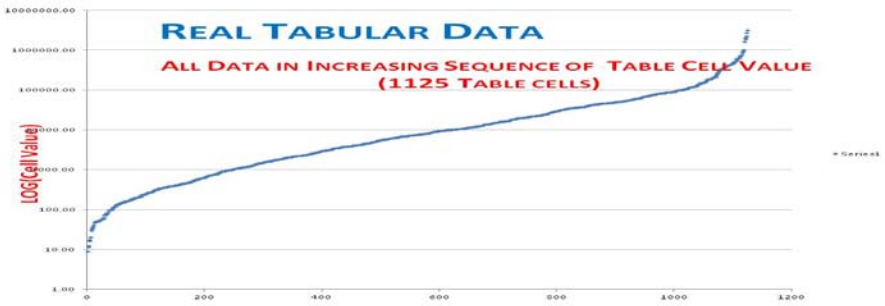
Fig.1.



[7] Input data used in this paper and optimum solutions (courtesy of Prof. Jordi Castro) are available from http://www-eio.upc.es/~jcastro/data/dandekar_2012.zip .

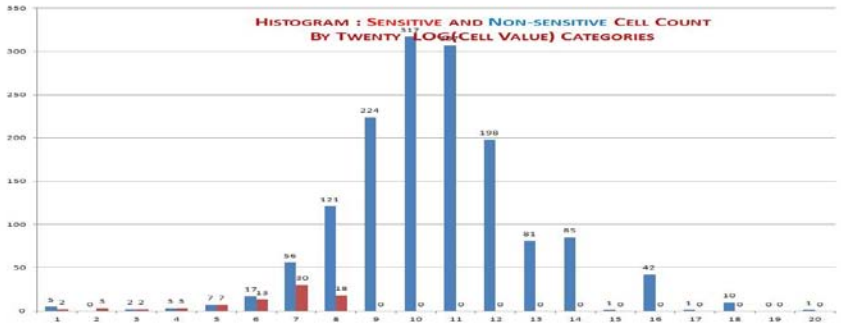
[8] Other synthetic tabular data generators which bypass generating simulated micro data records and directly populate tabular data cells are not capable of recreating complexities of real life multi-dimensional table structures and therefore are unsuitable for SDL research.

Fig. 2.



The histogram in Fig 3 uses the logarithmic scale for the table cell values and summarizes the distributions of sensitive and non-sensitive cells in the simulated table structure by dividing the range of table cell values in twenty equal-size intervals. Sensitive cells are clustered to the left of the non-sensitive cell distribution.

Fig. 3.



By using a distance measure, we have used two different cost functions of a) the cell value (referred to as L1 norm small) and b) the reciprocal of the cell value (referred to as L1 norm large) to evaluate the error distribution properties of the CTA model. The first cost function targets relatively small value table cells for adjustments, while the second cost function targets relatively large value table cells for adjustments. The near optimum solutions of the CTA model are courtesy of Prof. Jordi Castro[9]. The overall performance statistics from these two model runs are summarized in Table 2 by using eleven different percent error distribution categories. In both the options the majority of non-sensitive cells are unchanged[10]. Both options create Laplacian error distribution which is typical of the L1-norm linear regression model.

To demonstrate how the location of outliers changes depending on which error measure we use, in Figures 4 to 7 we have shown a synthetic non-sensitive data error distribution by using a "distance" measure typically used by the linear programming applications and by using a "percent error" measure proposed earlier in this paper to estimate the percent error associated with synthetic tabular data. For L1small and L1large

[9] J. A. González, J. Castro, A heuristic block coordinate descent approach for controlled tabular adjustment, Computers & Operations Research, 38 (2011) 1826-1835

[10] To further increase the total number of unchanged non-sensitive cells, Dandekar2001 paper demonstrates a simple iterative refinement of LP solution procedure in the section 5 of the paper.

options, by using a distance measure the vertical axis range is from -150 units to +150 units. The percent cell value change vertical axis is much larger for the L1small option (from -20% to +20%), as compared to that for the L1 Large option (from -4% to +5%).

Table 2.

OVERALL PERFORMANCE STATISTICS

L1 NORM LARGE 1088 NO CHNG

% From	% To	Non-Sens	Sensitive
.00 –	.00	1088	0
.00 –	.10	170	0
.10 –	.50	132	1
.50 –	1.00	57	0
1.00 –	1.50	9	0
1.50 –	2.00	10	1
2.00 –	5.00	12	50
5.00 –	10.00	0	26
10.00 –	15.00	0	0
15.00 –	30.00	0	0
30.00 –	100.00	0	0

L1 NORM SMALL 1233 NO CHNG

% From	% To	Non-Sens	Sensitive
.00 –	.00	1233	0
.00 –	.10	18	0
.10 –	.50	59	0
.50 –	1.00	49	0
1.00 –	1.50	33	0
1.50 –	2.00	20	0
2.00 –	5.00	47	36
5.00 –	10.00	13	29
10.00 –	15.00	3	6
15.00 –	30.00	1	4
30.00 –	100.00	2	3

Fig. 4.

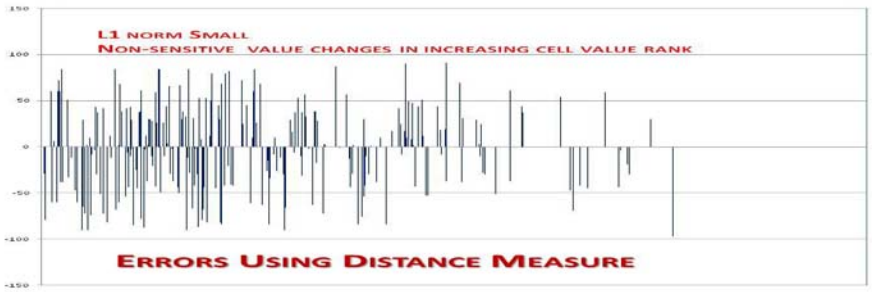


Fig. 5.

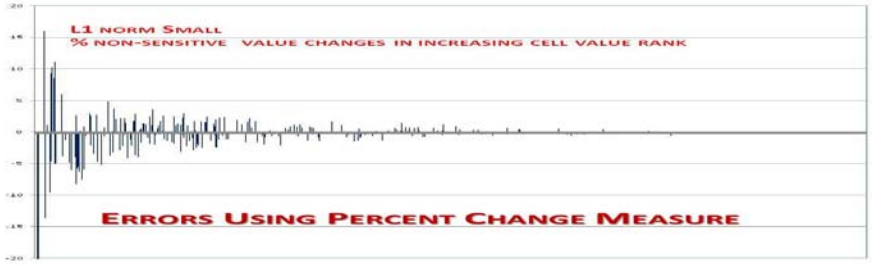


Fig. 6.

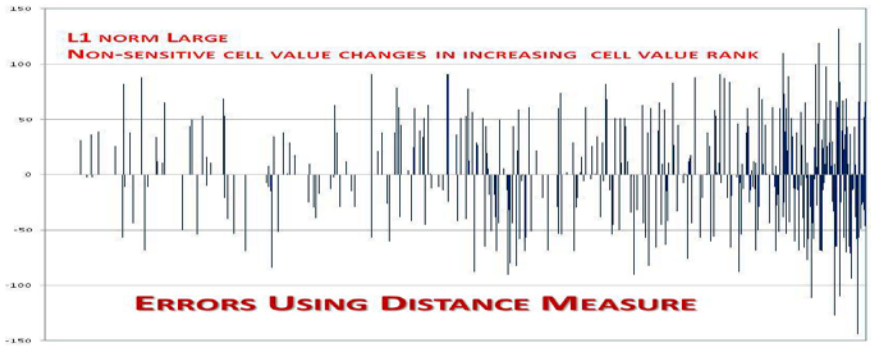
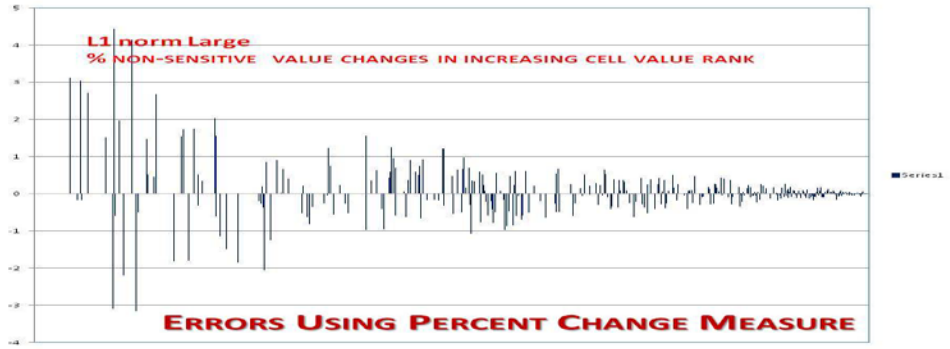


Fig. 7.



To demonstrate the effects of table cell value changes resulting from the CTA procedure on the overall cell distribution, in Table 3 we summarize before and after cell value distributions separately for sensitive and non-sensitive cells by using both the L1 norm small and the L1 norm large criteria. ***The CTA procedure trivially affects the relative cell distributions.***

Table 3.

L1small.fnl distribution before

From	To	Nonsensitive	Sensitive
1. -	2.	0.	0.
2. -	5.	0.	0.
5. -	10.	0.	0.
10. -	21.	1.	2.
21. -	44.	4.	2.
44. -	93.	2.	2.
93. -	198.	2.	3.
198. -	422.	9.	9.
422. -	899.	34.	25.
899. -	1914.	99.	32.
1914. -	4076.	257.	3.
4076. -	8678.	401.	0.
8678. -	18476.	356.	0.
18476. -	39337.	156.	0.
39337. -	83753.	102.	0.
83753. -	178320.	17.	0.
178320. -	379665.	26.	0.
379665. -	808351.	11.	0.
808351. -	1721075.	0.	0.
1721075. -	3664373.	1.	0.

L1small.fnl distribution after

From	To	Nonsensitive	Sensitive
1. -	2.	0.	0.
2. -	5.	0.	0.
5. -	10.	1.	0.
10. -	21.	1.	1.
21. -	44.	3.	3.
44. -	93.	1.	3.
93. -	198.	2.	2.
198. -	422.	9.	8.
422. -	899.	33.	22.
899. -	1914.	101.	34.
1914. -	4076.	257.	5.
4076. -	8678.	400.	0.
8678. -	18476.	356.	0.
18476. -	39337.	156.	0.
39337. -	83753.	102.	0.
83753. -	178320.	17.	0.
178320. -	379665.	26.	0.
379665. -	808351.	11.	0.
808351. -	1721075.	0.	0.
1721075. -	3664373.	1.	0.

L1large.fnl distribution before

From	To	Nonsensitive	Sensitive
1. -	2.	0.	0.
2. -	5.	0.	0.
5. -	10.	0.	0.
10. -	21.	1.	2.
21. -	44.	4.	2.
44. -	93.	2.	2.
93. -	198.	2.	3.
198. -	422.	9.	9.
422. -	899.	34.	25.
899. -	1914.	99.	32.
1914. -	4076.	257.	3.
4076. -	8678.	401.	0.
8678. -	18476.	356.	0.
18476. -	39337.	156.	0.
39337. -	83753.	102.	0.
83753. -	178320.	17.	0.
178320. -	379665.	26.	0.
379665. -	808351.	11.	0.
808351. -	1721075.	0.	0.
1721075. -	3664373.	1.	0.

L1large.fnl distribution after

From	To	Nonsensitive	Sensitive
1. -	2.	0.	0.
2. -	5.	0.	0.
5. -	10.	0.	0.
10. -	21.	1.	2.
21. -	44.	4.	2.
44. -	93.	2.	2.
93. -	198.	2.	4.
198. -	422.	9.	7.
422. -	899.	34.	25.
899. -	1914.	98.	33.
1914. -	4076.	258.	3.
4076. -	8678.	403.	0.
8678. -	18476.	354.	0.
18476. -	39337.	156.	0.
39337. -	83753.	102.	0.
83753. -	178320.	17.	0.
178320. -	379665.	26.	0.
379665. -	808351.	11.	0.
808351. -	1721075.	0.	0.
1721075. -	3664373.	1.	0.

8 Conclusion

We have demonstrated that the CTA procedure to generate synthetic tabular data, as originally proposed (Dandekar2001), is in fact a valid statistical procedure and utilizes a least absolute deviation linear regression model to achieve that goal. The model creates a error distribution with a relatively high peak and a rapid decay function. Due to highly skewed distribution of tabular data cells, ***Statistical measures advocated by QP-CTA are ineffective in further improving the overall data quality.*** The near optimum solutions used in this paper are possible only for problems with relatively few variables. In practice relatively large problems need to be solved by using heuristic procedure similar to that proposed in the Dandekar2001. In Appendix B we have summarized the outcome from the Dandekar2001 heuristic to allow comparisons with MILP solutions used in the paper.

References:

1. Dandekar R. A. (2001) "Synthetic Tabular Data: A Better Alternative To Complementary Data Suppression - Original Manuscript Dated December 2001". Energy Information Administration, U. S. Department of Energy. Also available from CENEX-SDC Project International Conference, PSD2006, Rome, Italy, December 13-15, 2006, Companion CD Proceedings ISBN: 84-690-2100-1.
2. 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.
3. 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)
4. 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.
5. 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)
6. 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)
7. 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.
8. Dandekar Ramesh A. (2009), "Incorporating Quality Measures in Tabular Data Protected by Perturbation Methods", Presented at FCSM2009 Conference.
9. 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.

10. 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.

Appendix A

From: Dandekar, Ramesh
To: aj.hundepool@cbs.nl; singh-avi@norc.org; Kelly@OptTek.com; tambjea@statcan.ca; jerry@stat.duke.edu; James.T.Fagan@Census.gov; John.Abowd@cornell.edu; jcastro@eio.upc.es; Josep Domingo (Business Fax); JJSalaza@ull.es; KrishM@uky.edu; Laura.Zayatz@census.gov; LWLG@CBS.nl; franconi@istat.it; MSRBCEL@fs1.ec.man.ac.uk; fisch@dei.unipd.it; Meena Khare (meena.khare@cdc.hhs.gov); Mcohen@nas.edu; n.shlomo@soton.ac.uk; Neil Russell (Neil.Russell@samhsa.hhs.gov); paul.b.massell@census.gov; sarathy@okstate.edu; rittle@umich.edu; Sarah Giessing (sarah.giessing@destatis.de); fienberg@stat.cmu.edu; Cohen, Stephen H.; Stephen Roehrig; tommy.wright@census.gov; teraghu@umich.edu; william.e.winkler@census.gov; yves.thibaudeau@census.gov; "karr@niss.org"
Subject: BY USING 1959 Harvey M Wagner paper it is ease to demonstrate that the CTA LP model is linear regression model
Date: Monday, April 23, 2012 9:09:30 AM
Attachments: WagnerLinear.pdf

Hi Folks

BY USING 1959 Harvey M Wagner paper, " Linear Programming Techniques for Regression Analysis", It is straight forward to demonstrate that the LP-based CTA formulation to generate synthetic tabular data is, in essence, a concurrent (simultaneous) execution of multiple least absolute deviation linear regression model(s) performed in such a way that they all meet the equality constraints imbedded in the CTA formulation.

How do you check that out?

Step 1: use Harvey M. Wagner formulation on a univariate error correction model specification on Y = X + error

Step 2: Apply this formulation separately to each table cell arranged as a column vector in the equality matrix m(delta)=0 in the CTA formulation

Step 3: Observe that each column in the CTA formulation is, in essence, a separate independent representation of the Wagner format error model for each table cell, arranged in such a way that all errors terms horizontally (row wise) satisfy the equality constraints associated with the CTA table structure.

IN SUMMARY, CTA MODEL IS CONDITONAL, CONCURRENT EXCECUTION OF LINEAR LEAST ABSOLUTE DEVIATION REGRESSION MODELS AIMED AT ACHIEVING EQUALITY CONSTRAINTS ASSOCIATED WITH CTA ERROR TERMS. THE CONDITIONAL ASPECTS ARE RELATED TO (1)STATISTICAL DISCLOSURE CONTROL RELATED BOUND AND (2) QUALITY RELATED UPPER TABLE CELL BOUNDS.

- RAMESH

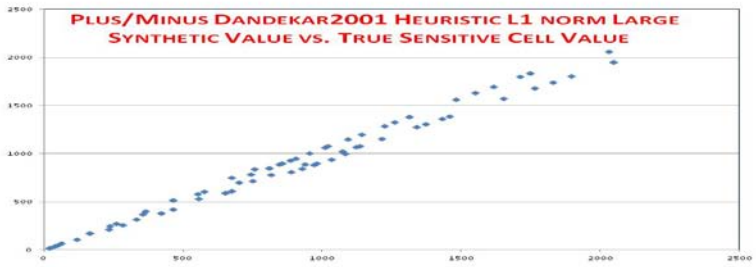
Appendix B

The summary outcome from Dandekar2001 plus/minus heuristic by using L1 norm large option, which is also available from the Dandekar2007 paper, is as follows:

PLUS/MINUS DANDEKAR2001 HEURISTIC
1081 NO CHNG

% From	% To	Non-Sens	Sensitive
.00 –	.00	1081	0
.00 –	.10	154	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

Fig. 8.



Figures 9 and 10 show the adjusted vs. true sensitive cell values for both L1 small and L1 large options used in this paper. For the L1 small option more of the larger sensitive cells appear to have been fixed at the lower bound. For the L1 Large option, multiple clusters of sensitive cells are changed in the same up or down direction.

Fig. 9.

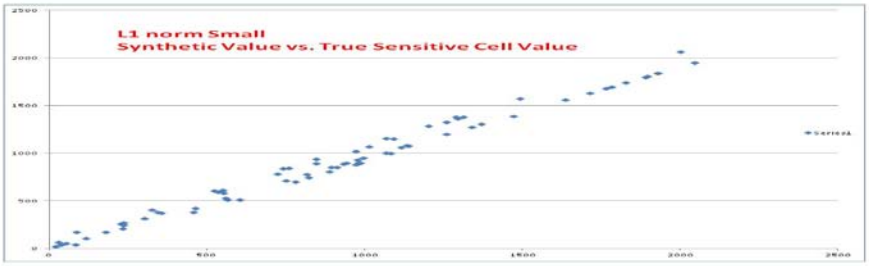
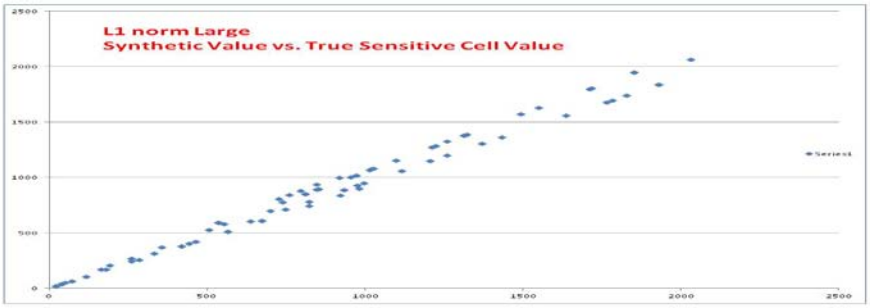


Fig. 10. .



Appendix C Presentation Slides

Statistical Basis of Controlled Tabular Adjustments to Generate Synthetic Tabular Data

[In]appropriate Use of Statistical Measures in [The Name of] Balancing Data Quality and Confidentiality of Tabular Format Magnitude Data

Joint Statistical Meetings 2012
Ramesh A. Dandekar , Mathematical Statistician
July 29, 2012, San Diego, California

Statement of Problem

Quality Preserving Controlled Tabular Adjustment (QPCTA) has been promoted to the statistical community since 2003 as a better alternative to the original CTA method proposed in Dandekar2001 Department of Energy manuscript

QPCTA uses arithmetic mean, variance, Pearson correlation to measure overall quality under UNREALISTIC assumption that tabular data cells follow a normal distribution

Current research in USA and in Europe on CTA does not take in to account the fact that relative change in table cell value (percent change) determines table cell quality and not change in table cell value by itself

Finding optimum or near optimum CTA solution is not essential to achieve overall tabular data quality

Presentation Outline

- First part of the presentation connects statistics and related operations research methods
  - A. Operations research tools for statistical applications
  - B. Unnecessary emphasis on optimality and reducing duality gap
  - C. Mostly ignores statistical aspects
- Second part addresses statistical aspects of synthetic tabular data deserving attention

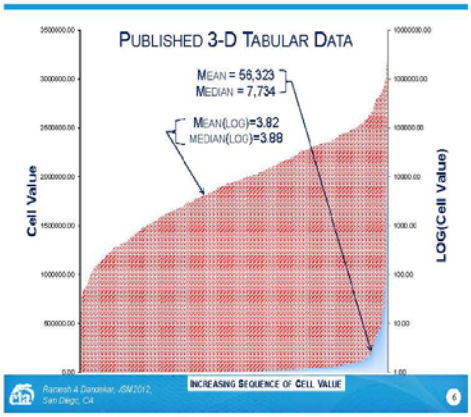
Connection of Mean and Median to Optimization Technique**

- Mean is the point that minimizes the sum of the squared deviations between the data points and mean value itself [Mean is estimated by using **least squares measure**]
- Median is the point that minimizes the sum of the absolute values of the difference between each data point and median itself. [Median is estimated by using **least absolute deviation measure**]

"Vanderbei, Robert, "Linear Programming Foundations and Extensions"

Distributional Considerations During Optimization

- For a symmetric distribution the mean and median are the same. As a result, the outcome from least square measures is similar to least absolute deviation measures
- For a skewed distribution the median is a more useful representation of central tendency than the mean. Consequently, the outcome from a least absolute deviation is a more appropriate and robust measure than the outcome from a least square measure



In regression problems alternative criteria of "best fit" to least squares are least absolute deviations and least maximum deviations. In this paper it is noted that linear programming techniques may be employed to solve the latter two problems. In particular, if the linear regression relation contains p parameters, minimizing the sum of the absolute value of the "vertical" deviations from the regression line is shown to reduce to a p equation linear programming model with bounded variables; and fitting by the Chebyshev criterion is exhibited to lead to a standard-form $p+1$ equation linear programming model.

Harvey M. Wagner, "Linear Programming Techniques for Regression Analysis", Journal of American Statistical Association, Vol 54, No 285 (March 1959), pp 206-212

FOR EACH TABLE CELL USES ERROR CORRECTION MODEL
$Y_{ESTIMATE} = X_{TRUE} + ERROR$
BY USING LEAST ABSOLUTE DEVIATION [L1 NORM] METHOD

1. MINIMIZING THE SUM OF ABSOLUTE DEVIATIONS

Let x_{ij}, $i=1, 2, \cdots, k$, and $j=1, 2, \cdots, p$, denote a set of k observational measurements on p "independent" variables, and y_i, $i=1, 2, \cdots, k$, denote the associated measurement on the "dependent" variable. Note that in the case of curvilinear regression, we may have $x_{ij}=z_j^2$, or $x_{ij}=\log z_{ij}$, or $x_{ij}=w_i z_{ij}$, etc. We wish to find regression coefficients b_j that

$$\text{Minimize}_{b_j} \sum_i \left| \sum_j x_{ij} b_j - y_i \right| . \quad (6)$$

Using the reduction in Charnes, Cooper, and Ferguson [1], the problem (6) is transformed into

$$\text{Minimize} \sum_i v_i + \sum_k w_k \quad (7)$$

To Transform The Objective Function To a Percent Change Measure from a Distance Measure

1) Simultaneously multiply and divide the cost function by related table cell value

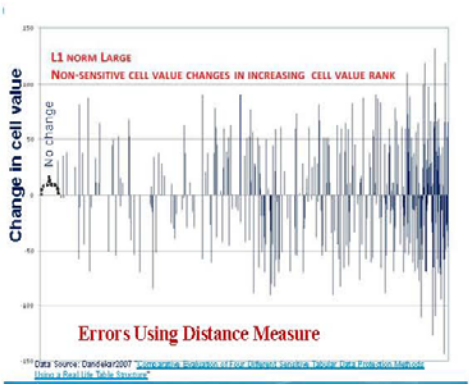
2) Associate the cell value in the denominator of the cost function with the variable in the objective function

EQUIVALENT COST FUNCTION MEASURES

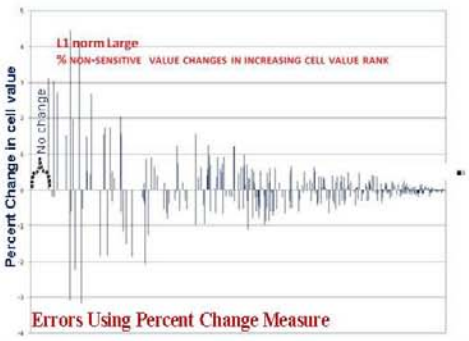
DISTANCE MEASURE	PERCENT CHANGE MEASURE
CONSTANT	CELL VALUE
CELL VALUE	CELL VALUE * CELL VALUE = (CELL VALUE)^2
LOG (CELL VALUE)	LOG (CELL VALUE) * CELL VALUE
1 / CELL VALUE	CONSTANT
1 / (CELL VALUE)^2	1 / CELL VALUE
[LOG(CELL VALUE)] / CELL VALUE	LOG(CELL VALUE)

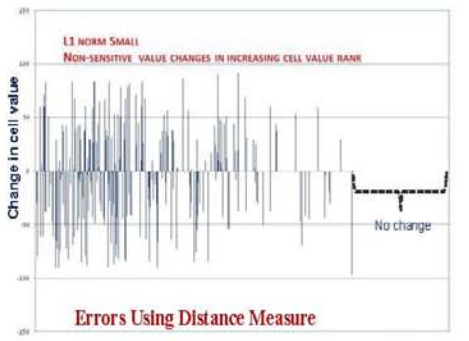
Location of Potential Outliers

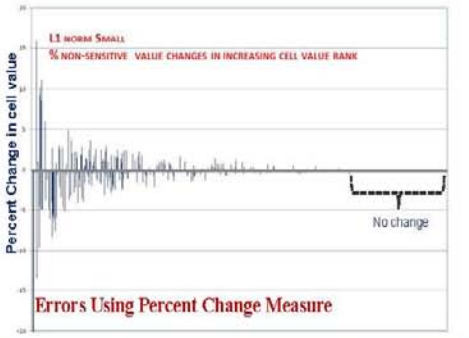
- Based on the distance measure, potential outliers are towards larger cells (used in the name of balancing data quality by QPCTA method)
- Based on the percent change measure, potential outliers are mostly in the region of smaller cells (could be used to further improve data quality)

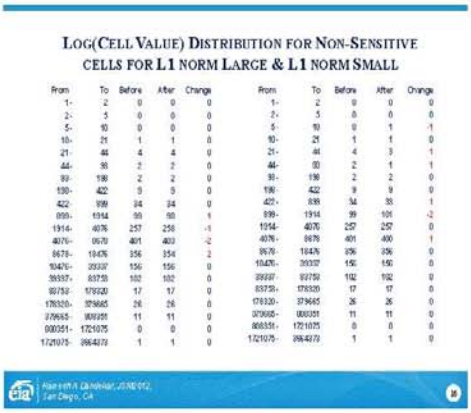


Data Source: Dandekar2007 "Comparative Evaluation of Four Different Sensitive Tabular Data Protection Methods Using a Real Life Table Structure"









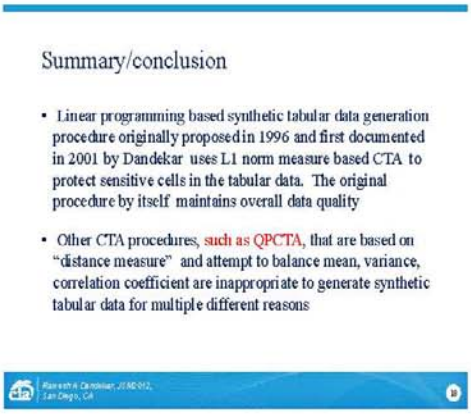
LOG(CELL VALUE) DISTRIBUTION FOR NON-SENSITIVE CELLS FOR L1 NORM LARGE & L1 NORM SMALL

From	To	Before	After	Change	From	To	Before	After	Change
1-	2	0	0	0	1-	2	0	0	0
2-	5	0	0	0	2-	5	0	0	0
5-	10	0	0	0	5-	10	0	1	-1
10-	21	1	1	0	10-	21	1	1	0
21-	44	4	4	0	21-	44	4	3	1
44-	98	2	2	0	44-	93	2	1	1
99-	198	2	2	0	99-	198	2	2	0
199-	422	9	9	0	199-	422	9	9	0
422-	899	34	34	0	427-	899	34	33	1
899-	1914	99	98	1	899-	1914	99	101	-2
1914-	4076	257	258	-1	1914-	4076	257	257	0
4076-	8679	401	403	-2	4076-	8679	401	400	1
8679-	18476	356	354	2	8679-	18476	356	356	0
18476-	39337	156	156	0	18476-	39337	156	156	0
39337-	83758	102	102	0	39337-	83758	102	102	0
83758-	178320	17	17	0	83758-	178320	17	17	0
178320-	379665	26	26	0	178320-	379665	26	26	0
379665-	808351	11	11	0	379665-	808351	11	11	0
808351-	1721075	0	0	0	808351-	1721075	0	0	0
1721075-	3664079	1	1	0	1721075-	3664079	1	1	0

Overall statistics non-sensitive cells

	Sum(abs(chng))	unchanged	changed	Aver over changed	Aver over all
L1small	9,712	1,233	323-78=245	30.07	6.24
L1Larg	15,096	1,088	468-78=390	32.25	9.7

	L1 Small All Cells	L1 Large All Cells
Correlation Coef	1.0	1.0
Mean True Values	22068.64	22068.64
Mean Adjusted Values	22068.64	22068.54



Summary/conclusion

- Linear programming based synthetic tabular data generation procedure originally proposed in 1996 and first documented in 2001 by Dandekar uses L1 norm measure based CTA to protect sensitive cells in the tabular data. The original procedure by itself maintains overall data quality
- Other CTA procedures, such as QPCTA, that are based on "distance measure" and attempt to balance mean, variance, correlation coefficient are inappropriate to generate synthetic tabular data for multiple different reasons