Working Paper No. 39
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 limitation

BUREAU OF TRANSPORTATION STATISTICS' PROTOTYPE DISCLOSURE LIMITATION SOFTWARE FOR COMPLEX TABULAR DATA

Contributed Paper

Submitted by the U.S. Department of Transportation, Bureau of Transportation Statistics
and
OptTek Systems, Inc. (United States)[1]

Abstract: The U.S. Department of Transportation, Bureau of Transportation Statistics (BTS), initiated a project to develop and implement a new, state-of-the-art statistical disclosure limitation (SDL) method for multi-dimensional (up to five) hierarchical tables. The goals of the project were to increase the amount of information released (compared to cell suppression), protect the identity of the data providers, and maintain the statistical properties of the tables. After reviewing a wide variety of SDL methods for tabular data, the project team selected the Synthetic Tabular Data (STD) method proposed by Dandekar and Cox (2002). STD was subsequently renamed Controlled Tabular Adjustment or CTA (Cox and Dandekar 2002). CTA is a new SDL method for tabular data that can be used in lieu of cell suppression. CTA imputes safe values for sensitive tabular cells and then adjusts other cells to restore additivity to the table. CTA evaluated well for BTS requirements and goals. This paper briefly discusses the research process and development of an efficient algorithm based on the CTA method and Tabu Search (Glover and Laguna 1997) that allows for processing of multi-dimensional tables with hundreds of thousands of entries designed for use in a static system. (A static system is where the only tables published are those chosen by a National Statistical Offices (NSO), and the microdata that created the tables would not be released.) BTS is the first NSO to use a modified version of the CTA method bundled with Tabu Search for implementation into prototype computer software for testing and demonstration. The software user interface, software functions and options are described, and examples of processed tables using NSO tabular data are provided. In general, the modified CTA method implemented preserved the statistical properties of the tabular data when original tables were compared to the modified (protected) tables. The paper concludes with suggestions for future software redesigns and research.

I. INTRODUCTION

1. National Statistical Offices (NSO) develop confidentiality policy to protect individually identifiable information that is collected for statistical purposes. The United States Department of Transportation, Bureau of Transportation Statistics (BTS) has drafted a confidentiality policy based on its legislative mandate (49 U.S.C. 111(i)) to protect individually identifiable information. The BTS

[1] Prepared by J. Neil Russell (Neil.Russell@bts.gov) and James P. Kelly (Kelly@opttek.com)

confidentiality policy recommends methods to maintain confidentiality of individually identifiable information for tabular data and micro data. Because SDL methods are still evolving, BTS decided to sponsor research to develop new ways that could increase access to data while protecting confidential data at the same time. The goals of the research project were to 1) develop a new method of protecting tabular data, 2) increase access to data (when compared to complementary cell suppression), and 3) ensure that application of this new method would not substantially affect the quality (statistical properties) of the data.

2. The goals of this project were an outgrowth of BTS' experience with releasing only tabular data from certain surveys. One example where only tabular data are released is the Commodity Flow Survey (CFS). The CFS has been designed to provide data on the flow of goods and materials by mode of transport. The survey was conducted in 1993, 1997, and 2002. The U.S. Bureau of Census (BOC) conducted the data collection on behalf of BTS. CFS data releases are subject to the strict requirements of BOC's confidentiality legislation (see 13 U.S.C. 9 and 13 U.S.C. 214). In this case, only tabular data are released to the public. The CFS tables of data have been protected through use of complementary cell suppression, and as a result, many tables have a great deal of missing information. Data users of the CFS find that the quantity of missing information hinders analysis of the data, for much of the needed information is not available. To this end, BTS initiated research to explore alternatives to complementary cell suppression.

II. PROJECT GOALS AND NEW SDL METHOD FOR COMPLEX TABULAR DATA

3. The research project first started by reviewing known SDL methods for tabular data, and then assessing their utility for meeting the goals of the project (Russell et. al. 2002). The literature review identified the Synthetic Tabular Data (STD) method (proposed by Dandekar and Cox 2002) as a likely approach that would fulfill the project goals. STD was subsequently renamed Controlled Tabular Adjustment or CTA (Cox and Dandekar 2002). The CTA method generally works in two steps: 1) identify sensitive cells and replace those values with imputed values from a calculated range of projection, and 2) adjust non-sensitive cells a minimal amount to foster table additivity while not significantly changing the statistical properties of the table.

4. CTA was chosen for implementation into software because it evaluated well when compared to the project goals. For example, complementary cell suppression does protect respondent identity, but it usually blanks out large amounts of data in multi-dimensional tables. Hence, complementary cell suppression would not meet the BTS goal of "increased access to data." The CTA method produces fully populated tables, unlike complementary cell suppression, and achieves the goal of increased access to data.

5. The related goal of protecting data is also met, for the CTA method focuses on protecting those cells in a table that could be used to ascertain the identity of data providers. Values for these sensitive cells are imputed at a "sufficient distance" from the true value in order to maintain confidentiality. Sensitive cells are not identified, and with fully populated tables, the intruder is hindered in attempts to re-identify data providers. This, then, achieves the BTS goal of "protecting the identity of data providers."

6. The third goal of the project was to minimize changes (i.e.: minimize information loss) to the statistical properties of the tables once the method was applied. SDL methods should have little or no impact on statistical data uses. However, it is impossible to anticipate every possible use of data. Thus, it was proposed that measures of information loss, if minimized, should produce modified tabular data that will produce statistical analyses very similar to the original, unprotected data. To assess and understand

information loss, correlation analysis was planned as one reasonable measure. In particular, consider *x* as the elements of an unprocessed table, and *x'* as elements of a protected table:

- Compute the correlation (r_{all}) of *x* to *x'* for all tabular entries.
- Compute the correlation ($r_{sensitive}$) of *x* to *x'* for only the sensitive cells.
- Compute the correlation ($r_{nonsensitive}$) of *x* to *x'* for only the nonsensitive cells.

A processed table that produces $r^2_{all} \sim 1$, $r^2_{sensitive} \sim 1$, and $r^2_{nonsensitive} \sim 1$ is one that minimizes information loss. These proxies for information loss were used to assess CTA and to enable characterization of changes to the statistical properties of tables. The project also adopted other information loss measures as proposed by Domingo-Ferrer and Toora (2001).

III. RESEARCH PROJECT PHASES

7. In order to implement CTA for testing and use, the research project was sectioned into four phases. Each phase is briefly described, with this paper mainly focusing on phase 3.

- In phase 1, CTA was implemented using integer programming models and solvers, fashioned to be similar to how Dandekar and Cox (2002) originally proposed it. This allowed optimal solutions to be found for small to medium-sized problems. The optimal solutions generated using exact algorithms are valuable in testing and evaluating the performance of the heuristic algorithms developed in phase 2.
- Phase 2 is where CTA was modified through implementation with Tabu Search (Glover and Laguna 1997) heuristics. Here, the heuristics algorithms were developed utilizing state-of-the-art adaptive memory techniques to seek near optimal solutions. More information is available on this phase of research (see BTS 2002b).
- In phase 3, the heuristic algorithms developed in phase 2 were integrated into prototype software designed to integrate with SPSS statistical software. The prototype software was designed with a graphical user interface for control.
- And in phase 4, the prototype software was tested using simulated and real data. The goal was to fine tune the algorithm as well as quantify its effectiveness in minimizing information loss.

IV. BTS PROTOTYPE SDL SOFTWARE FOR TABULAR DATA

8. With the minimization of information loss as a project goal, research and experimentation settled on a non-linear objective function as the heuristic algorithm to be implemented as developed in phase 2 of the project. In general, the heuristic seeks to minimize the *Mean Variation* divided by the *Correlation* between the modified and original sensitive cells raised to the *p* power. Specifically, let *X* be the original, unprocessed table with elements *x*, *X'* is the protected table with elements *x'*, and *n* is the number of tabular elements.

Where,

$$\text{Mean Variation} = \sum_{x \in X,\, x' \in X'} (|x - x'| / |x|) \;/\; n$$

And,

$$\text{Correlation Coefficient} = \text{Corr}(x \text{ to } x')$$

Hence,

$$\text{Minimize} \quad \text{Mean Variation} \; / \; (\text{Correlation Coefficient})^{p}$$

This nonlinear objective function encourages this CTA method to seek solutions with small relative changes and correlations (compared to the original data) that are approximately equal to 1. Preliminary analysis has indicated that the correlation raised to the tenth power has the effect of increasing the correlation with little impact on the mean variation.

9. This general heuristic algorithm was specified using object-oriented analysis and object-oriented design and written using the "C" and "C++" programming languages. Both the design and programming code went through a review process for quality assurance. They were developed as a separate library of algorithms that are available as a Dynamic Link Library (DLL) for Microsoft Windows. This interface was also designed to integrated into other Windows Applications besides SPSS.

10. Processing tables in the BTS prototype software follows these steps:

- In SPSS, create desired table, and then save the programming code (syntax file) used to generate the table.
- The BTS prototype software is started from the SPSS File menu. Inputs to the software are the SPSS data file and syntax file used to generate the original table.
- Specify the sensitivity rule to be applied to protect the data.
- When the modified CTA algorithm is executed, the microdata is read in and processed based on specifications in the syntax file and the sensitivity rule(s) selected. A new SPSS data file is created.
- The modified (protected) tables are generated by running the original syntax file against the new data file.

11. The software interface is given in Figure 1. The features of the software interface and their uses are described as:

- Input Files – the software takes an SPSS data file (.sav extension) and syntax file (.sps or .txt extension) as inputs. The Browse button is used to locate the data file and syntax file. The data View button allows viewing the data file. The More button will read in additional records. The syntax View button displays the contents of the syntax file.
- Output File - The software creates a SPSS data file as output that is only used for regenerating the final tables. The software never modifies the input data file or input syntax file. Also, structural zeros are never modified in final tables. The Rename button allows one to change the file name and location of the output data file. The output data file contains only data for the fields used to regenerate the final modified table. The output data file will not have the same number of records as the input file. For tables using count data, the output file will have the number of cases needed to regenerate the modified counts. For magnitude table data, the output file will have a case for each published cell where the summed field will reflect the new value assigned by the modified CTA method.
- Log File - A log file is automatically generated each time the Solve button is selected. The log file resides in the same directory as the syntax file and has the same name as the syntax file but with the extension .log. The log file can be viewed using any ASCII text viewing software. The log file lists the input files, output files, sensitivity rules, table parameters and information loss statistics. The option is given for appending or overwriting an existing log file.
- Sensitivity Rules – This is where one specifies which rule or rules to use to protect the tabular data. If multiple rules are applied to data, the rules are cumulative. If more than one rule is applicable to a cell, the protection range is set to the least restrictive value. If the syntax file calls

for count data, only the count data sensitivity rules will be active. If the syntax file calls for magnitude data, the magnitude data sensitivity rules will be active. The sensitivity rules to choose from are the Number Rule, Sum Rule, Percent Rule, Prior Posterior Rule, or Dominance Rule. More information is available on these rules and how the protection ranges are calculated for the sensitive cells (see BTS 2002a).

- Deviation for Non-Sensitive Cells - The deviation for non-sensitive cells defines how much non-sensitive data cells and non-sensitive sums are allowed to vary. The deviation is calculated as a percentage of the original cell value. The default value is 20%.
- Discourage Sum Changes – Checking this options discourages non-sensitive sums from changing from original values. The default is to treat non-sensitive sums the same as non-sensitive data cells.
- Output Statistics - When the software has completed solving the problem, sensitivity rules, table parameters, and information loss statistics will be displayed in the interface. The information loss statistics compare the original table to the protected table. All statistics displayed on the software interface are written to the log file.
- Generating the Report - The Generate Report button will regenerate the tables using the syntax file and the new output data file. This button also invokes SPSS, and by using the SPSS File/Save interface, the protected tables can be saved.

V. RESULTS FROM USING BTS PROTOTYPE SOFTWARE

12. About 162 different tabular data test cases have been processed to date and information loss statistics are summarized for them in Table 1.

Table 1
Summary of Information Loss Statistics

Tabular Data Type	Range of Correlation		
	Sensitive Cells	Non-sensitive Cells	All Cells
Count	.63 to .99	.99 to 1.00	.99 to 1.00
Magnitude	.89 to .99	.99 to 1.00	.99 to 1.00

Source: BTS SDL Method for Tabular Data Research Project

These results provide some evidence that the modified CTA method implemented into BTS prototype software generally maintains the statistical properties of the original tables.

13. Figure 2 gives one example of a three-dimensional table using magnitude data. The sensitivity rule specifications, table parameters, and information loss statistics are given first. The original tables and modified (protected) tables are also reproduced in Figure 2.

VI. CONCLUSION

14. BTS was the first NSO to successfully implement a modified version of the CTA method bundled with Tabu Search into prototype software. The research to date points to CTA as an effective SDL method for tabular data. The goals of BTS' research project were met where a new SDL method was implemented that can protect tabular data, increase data access, and maintain the statistical properties of the original tables.

15. Some future work on the CTA method and software could include the following:

- Perform model-based optimization - One of the potential criticisms of CTA is that it may modify data in such a way as to adversely affect statistical models derived from the tabular data. To explore this issue, model-based optimization analyses should be performed. Specifically, the software and algorithms could be modified to preserve model integrity during the CTA process. Comparisons can be done to contrast model-based solutions with those derived from generic objective functions. Furthermore, the system can be expanded to allow users to perform this model-based optimization.
- Identify and analyze alternative objectives - The system should be extended to consider alternative objectives such as Chi Square, regression slope, or changes in variance. It is highly likely that certain types of tables could be better-processed using objectives that are related to the tabular data types. The resulting system would be more robust in its ability to better-process wide ranges of data.
- Develop additional interfaces - The current system has been interfaced to SPSS. Interfacing the system to other statistical software, such as SAS, would extend its potential use to a wider audience.
- Pursue more advanced formulations of Tabu Search - The current system utilizes a straightforward implementation of Tabu Search to search for the optimal solution. Research and software development into more advanced Tabu Search techniques would improve the effectiveness and efficiency of the system. The resulting system would produce tables that are closer to optimal. The exploration of better methods to minimize changes to sums would also be desirable.
- Analyze the efficacy of rule-based sensitive cell determination - The rules that determine sensitive cells and their allowable upper and lower bound values should be thoroughly analyzed. The system could be modified to allow different parameters or criteria for setting upper and lower bounds for sensitive cells. The system could also intelligently suggest appropriate rules to use and/or produce alternative solutions for user consideration.

VII. REFERENCES

Bureau of Transportation Statistics (BTS). "Review of the State of the Disclosure Limitation Methods for Tabular Data and Research Proposal". Task 1 Final Report. Washington, DC. 2002a.

Bureau of Transportation Statistics (BTS). "Disclosure Limitation Methods for Tabular Data". Task 2 Final Report. Washington, DC. 2002b.

Cox, Larry H. and Ramesh A. Dandekar. "A Disclosure Limitation Method for Tabular Data that Preserves Data Accuracy and Ease-of-use." Paper presented at the Federal Committee on Statistical Methodology Statistical Policy Conference. Bethesda, MD. 2002.

Dandekar, Ramesh A. and Larry H. Cox. "Synthetic Tabular Data - An Alternative to Complementary Cell Suppression." Unpublished Manuscript. 2002.

Domingo-Ferrer, Josep. and Vicenc Torra. "Disclosure Control Methods and Information Loss for Microdata", in *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, edited by Doyle, P., Lane, J., Theeuwes, J., and Zayatz, L., New York, North-Holland Elsevier, 2001.

Glover, Fred. and Manuel Laguna. *Tabu Search*, Kluwer Academic Publishers. 1997.

Russell, J. Neil, James P. Kelly, and Fred Glover. "The Bureau of Transportation Statistics' Statistical Disclosure Limitation Method for Tabular Data: A Review", 2002 Proceedings of the American Statistical Association, Government Statistics Section [CD-ROM], Alexandria, VA:American Statistical Association, 2002.

Figure 1

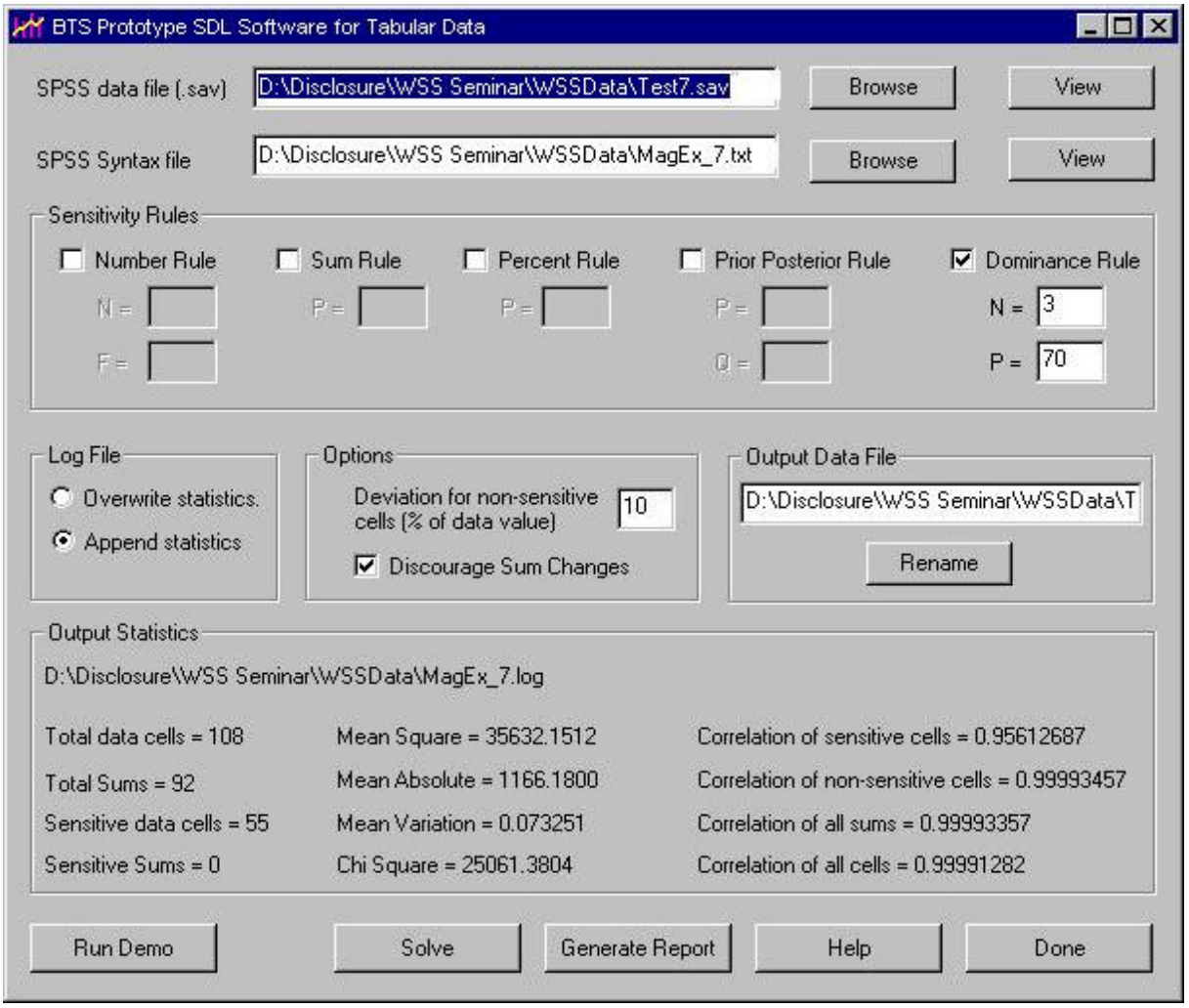

Figure 2
Output Statistics For CTA Processed Magnitude Tabular Data Example

Sensitivity Rules
Dominance Rule: N = 3, P = 70
Discourage Sum Changes
Deviation for non-sensitive cells = 10%

Table Parameters
Total data cells = 108
Total Sums = 92
Sensitive data cells = 55
Sensitive Sums = 0

Information Loss Statistics
Mean Square = 35632.1512
Mean Absolute = 1166.1800
Mean Variation = 0.073251
Chi Square = 25061.3804
Correlation of sensitive cells = 0.95612687
Correlation of non-sensitive cells = 0.99993457
Correlation of all sums = 0.99993357
Correlation of all cells = 0.99991282

Figure 2 - Continued
Original Magnitude Tabular Data Example

Season PL Summer

			Census Region				Table Total
			Northeast	Midwest	South	West	
			Sum	Sum	Sum	Sum	Sum
Year Built	LE 1899		4282.00	2756.00	2528.00	112.00	9678.00
	1900 - 1919		8787.00	10033.00	1804.00	2108.00	22732.00
	1920 - 1945		43598.00	36838.00	72330.00	4187.00	156953.00
	1946 - 1959		43635.00	63789.00	34667.00	18675.00	160766.00
	1960 - 1969		112099.00	46054.00	109190.00	50188.00	317531.00
	1970 - 1979		120863.00	95859.00	151607.00	95879.00	464208.00
	1980 - 1989		69362.00	82320.00	231686.00	79843.00	463211.00
	1990 - 1992		29382.00	25539.00	60894.00	23142.00	138957.00
	1993 - 1995		20746.00	21373.00	32475.00	16151.00	90745.00
Table Total			452754.00	384561.00	697181.00	290285.00	1824781.00

Season PL Winter

			Census Region				Table Total
			Northeast	Midwest	South	West	
			Sum	Sum	Sum	Sum	Sum
Year Built	LE 1899		1786.00	221.00	91.00	0	2098.00
	1900 - 1919		6207.00	4583.00	517.00	3650.00	14957.00
	1920 - 1945		10887.00	9631.00	7108.00	29078.00	56704.00
	1946 - 1959		6248.00	28885.00	20505.00	20152.00	75790.00
	1960 - 1969		12822.00	50839.00	64157.00	10906.00	138724.00
	1970 - 1979		41084.00	92134.00	47332.00	28600.00	209150.00
	1980 - 1989		43551.00	59880.00	121210.00	68515.00	293156.00
	1990 - 1992		9620.00	47917.00	24703.00	49430.00	131670.00
	1993 - 1995		1766.00	44009.00	8655.00	7751.00	62181.00
Table Total			133971.00	338099.00	294278.00	218082.00	984430.00

Season PL Sum & Win

			Census Region				Table Total
			Northeast	Midwest	South	West	
			Sum	Sum	Sum	Sum	Sum
Year Built	LE 1899		5.00	78.00	0	0	83.00
	1900 - 1919		128.00	538.00	306.00	95.00	1067.00
	1920 - 1945		574.00	1707.00	6466.00	671.00	9418.00
	1946 - 1959		141.00	7281.00	2848.00	50.00	10320.00
	1960 - 1969		1228.00	803.00	9201.00	401.00	11633.00
	1970 - 1979		49.00	5803.00	10031.00	2342.00	18225.00
	1980 - 1989		828.00	1199.00	8250.00	5983.00	16260.00
	1990 - 1992		311.00	10.00	948.00	0	1269.00
	1993 - 1995		159.00	447.00	1633.00	3141.00	5380.00
Table Total			3423.00	17866.00	39683.00	12683.00	73655.00

Table Total

			Census Region				Table Total
			Northeast	Midwest	South	West	
			Sum	Sum	Sum	Sum	Sum
Year Built	LE 1899		6073.00	3055.00	2619.00	112.00	11859.00
	1900 - 1919		15122.00	15154.00	2627.00	5853.00	38756.00
	1920 - 1945		55059.00	48176.00	85904.00	33936.00	223075.00
	1946 - 1959		50024.00	99955.00	58020.00	38877.00	246876.00
	1960 - 1969		126149.00	97696.00	182548.00	61495.00	467888.00
	1970 - 1979		161996.00	193796.00	208970.00	126821.00	691583.00
	1980 - 1989		113741.00	143399.00	361146.00	154341.00	772627.00
	1990 - 1992		39313.00	73466.00	86545.00	72572.00	271896.00
	1993 - 1995		22671.00	65829.00	42763.00	27043.00	158306.00
Table Total			590148.00	740526.00	1031142.00	521050.00	2882866.00

Figure 2 - Continued
CTA Processed Magnitude Tabular Data Example

Season PL Summer

			Census Region				Table Total
			Northeast	Midwest	South	West	
			Sum	Sum	Sum	Sum	Sum
Year Built	LE 1899		4509.00	2947.00	2098.00	146.00	9700.00
	1900 - 1919		8597.00	11232.00	1624.00	1906.00	23359.00
	1920 - 1945		44363.00	37779.00	71214.00	3768.00	157124.00
	1946 - 1959		43637.00	63819.00	34519.00	18674.00	160649.00
	1960 - 1969		112097.00	46112.00	105911.00	51370.00	315490.00
	1970 - 1979		120850.00	94231.00	153598.00	96111.00	464790.00
	1980 - 1989		67572.00	81623.00	233723.00	80293.00	463211.00
	1990 - 1992		29236.00	28093.00	60021.00	23544.00	140894.00
	1993 - 1995		17347.00	19236.00	33746.00	16072.00	86401.00
Table Total			448208.00	385072.00	696454.00	291884.00	1821618.00

Season PL Winter

			Census Region				Table Total
			Northeast	Midwest	South	West	
			Sum	Sum	Sum	Sum	Sum
Year Built	LE 1899		1399.00	260.00	118.00	.00	1777.00
	1900 - 1919		5025.00	4894.00	652.00	3823.00	14394.00
	1920 - 1945		10214.00	8861.00	6397.00	36159.00	61631.00
	1946 - 1959		6221.00	30720.00	20488.00	20143.00	77572.00
	1960 - 1969		13060.00	50721.00	67122.00	9815.00	140718.00
	1970 - 1979		41082.00	92134.00	47334.00	28600.00	209150.00
	1980 - 1989		45189.00	60229.00	119086.00	68518.00	293022.00
	1990 - 1992		9859.00	33922.00	25292.00	49430.00	118503.00
	1993 - 1995		1922.00	55316.00	7789.00	6892.00	71919.00
Table Total			133971.00	337057.00	294278.00	223380.00	988686.00

Season PL Sum & Win

			Census Region				Table Total
			Northeast	Midwest	South	West	
			Sum	Sum	Sum	Sum	Sum
Year Built	LE 1899		6.00	66.00	.00	.00	72.00
	1900 - 1919		90.00	467.00	398.00	124.00	1079.00
	1920 - 1945		482.00	1536.00	8293.00	470.00	10781.00
	1946 - 1959		166.00	5416.00	3013.00	60.00	8655.00
	1960 - 1969		992.00	863.00	9515.00	310.00	11680.00
	1970 - 1979		64.00	7431.00	8038.00	2110.00	17643.00
	1980 - 1989		980.00	1547.00	8388.00	5530.00	16445.00
	1990 - 1992		218.00	7.00	1232.00	.00	1457.00
	1993 - 1995		207.00	313.00	1228.00	4079.00	5827.00
Table Total			3205.00	17646.00	40105.00	12683.00	73639.00

Table Total

			Census Region				Table Total
			Northeast	Midwest	South	West	
			Sum	Sum	Sum	Sum	Sum
Year Built	LE 1899		5914.00	3273.00	2216.00	146.00	11549.00
	1900 - 1919		13712.00	16593.00	2674.00	5853.00	38832.00
	1920 - 1945		55059.00	48176.00	85904.00	40397.00	229536.00
	1946 - 1959		50024.00	99955.00	58020.00	38877.00	246876.00
	1960 - 1969		126149.00	97696.00	182548.00	61495.00	467888.00
	1970 - 1979		161996.00	193796.00	208970.00	126821.00	691583.00
	1980 - 1989		113741.00	143399.00	361197.00	154341.00	772678.00
	1990 - 1992		39313.00	62022.00	86545.00	72974.00	260854.00
	1993 - 1995		19476.00	74865.00	42763.00	27043.00	164147.00
Table Total			585384.00	739775.00	1030837.00	527947.00	2883943.00