Applying Cell Suppression to Inter-Related Tables

By Jacob Bournazian, Michael Salpeter, and Bin Zhang

Abstract

Cell suppression is the most common disclosure limitation method that the U.S. Energy Information Administration (EIA) applies to the aggregate statistics that it publicly releases. Complementary cell suppression involves withholding the publication of non-sensitive cells in a table in order to protect the cells that were identified as sensitive to revealing company level information. If several tables within the same information product are related due to a multi-dimensional table design, then the selection of complementary cells in each table becomes a more complicated computational problem. EIA has two different automated suppression programs that it applies to tabular data. Each program follows a different methodology. The data protection levels for a table vary as a result of applying different complementary cell suppression methodologies. This paper focuses on the situation when tables are related and compares two different automated approaches for applying cell suppression methodology.

Keywords: cell suppression, confidentiality, sensitivity, complementary suppression

Introduction

Cell suppression is a common method used by Federal statistical agencies to protect the confidentiality of information reported by individuals and businesses when publicly releasing tabular data. The objective is to protect the individual survey responses that are part of suppressed cells from being closely estimated by using the aggregate statistics shown in a table for non-suppressed cells.[1] The first step before applying any disclosure limitation method is to identify those cells in a table that are likely or "sensitive" to reveal reported values by a single company or individual. Various measures of sensitivity, defined as linear sensitivity rules, have been developed and are applied by statistical agencies.[2] The definitions and mathematical properties of these linear sensitivity measures and their ability to identify sensitive cells in a table have been analyzed.[3] The sensitivity rules that Federal statistical agencies apply are well documented in Statistical Working Paper No. 22.[4] The threshold levels that agencies set when applying those sensitivity rules are not publicly released.

For this research, the *p*-percent rule was used as the sensitivity rule for identifying sensitive cells in the table. The *p*-percent rule assumes that any respondent, that has its reported value included in an aggregate table cell value, can estimate the contribution of another respondent to within 100-percent of its value. This means that the estimating respondent knows that the other respondents' values are nonnegative and less than two times the actual value.[5] In applying this rule, the goal is that after releasing the aggregate tabular statistics, no respondent's value may be estimated more accurately than within *p* percent of the corresponding actual reported value.

Cells in a table that are identified by a sensitivity rule as sensitive to disclosing protected information are withheld from public release and are called primary suppressions. In order to safeguard the primary suppressions from being derived by subtraction from the published marginal totals and from being closely estimated with the constraints among the data in the table, additional non-sensitive cells within a table are also withheld. These additional non-sensitive cells that are selected and withheld from publication are called "complementary" suppressions. The selection of complementary cells to suppress in a table to protect primary suppressions is called a suppression pattern.

Software programs that identify primary cell suppressions and select complementary suppressions have been used by Federal statistical agencies since the mid-1970s.[6] These programs commonly use a linear programming method based on the hierarchical structure of the data. These programs can be used on both magnitude and frequency data tables. The selection of a suppression pattern for a table becomes problematic when table contains three or more dimensions or the table is linked or related to other published tables. This paper evaluates the performance of two automated suppression programs when tables are interrelated and contain three or more dimensions.

Research Approach

Motor gasoline wholesale volumes are reported by grade, formulation, sales type, and State by all refiners in the US. Aggregated data categorized by these four variables are published in Tables 39 and 40 of the *Petroleum Marketing Monthly (PMM)* publication.[7] These two tables are related (linked) three dimensional tables. One table shows refiner motor gasoline wholesale volumes by grade, sales type, and State; and the other table shows refiner motor gasoline wholesale volumes by formulation, sales type, and State. All marginal and grand totals are also included in the tables. The geography dimension has three hierarchical levels: three sub-regional totals; five regional totals; and the U.S. total. The categories in the grade dimension include Regular, Midgrade, Premium, and All Grades. The categories in the formulation dimension include Conventional, Reformulated, and All Formulations.[8] The categories in the sales type dimension include Dealer Tank Wagon (DTW), Rack, Bulk, and All Sales Types.[9]

Publishing the marginal and grand totals for wholesale sales made the interior wholesale table cells vulnerable for identifying the values of the suppressed cells. In 2008, the table design for these two tables was modified to delete the marginal wholesale totals column by sales type (all sales types) to reduce disclosure risk and reduce the amount of suppression needed in the table. The data in the three sales type categories of DTW, rack, and Bulk became independent of each other by not publishing the marginal column for total wholesale sales. Therefore, each table can be viewed as three independent tables displayed side by side, or three separate two dimensional tables, e.g. refiner motor gasoline DTW sales volumes by Grade and State.

Tables 39 and 40 in the *PMM* show gasoline sales by Grade of gasoline, by State and Formulation of gasoline by State. The two tables are linked because the column, “Sales of All Grades of Gasoline,” in the first table is identical to the column, “Sales of All Formulations of Gasoline” in the second table. If the space of all data items defined by all the intersections of Grade, Formulation, and State is viewed as a cube, then the data items in the two linked tables are two adjacent surfaces of the cube, with the common (identical) column as the shared edge.

This study is to compare two suppression programs that use different approaches to generate complementary cell suppression patterns in three dimensional tables that are linked. The DiAna software system uses a linear programming approach to identify complementary cells to suppress. The PMM (Petroleum Marketing Monthly) suppression program uses a non-linear programming approach.

The two programs were evaluated based on: the total number of complementary cells; and the corresponding total volumes that were suppressed to protect the primary suppressions; similarity of complementary suppression patterns generated by the two

different approaches; and the amount of disclosure risk in the suppression patterns from the two systems. Empty cells and cells with zero values were excluded as candidates for complementary suppression.

In order to compare the suppression patterns generated by the two programs, the list of primary suppression cells used by the two programs was the same. The *p*-percent rule was used with a value set at 5% for this project to define sensitive cells (primary suppressions). Data from the *PMM* for three months, March, July, and October 2009, were used for the comparison study in this paper. The total number of cells tested and the number of primary suppressed cells are shown below in Table 1.

Table 1. Counts of Primary Suppressions

	DTW		Rack		Bulk	
Month	Total	Primary	Total	Primary	Total	Primary
March	222	90	317	22	159	84
July	229	85	316	17	167	90
October	224	87	317	21	167	93

Among the three wholesale types of DTW, Rack, and Bulk, Rack sales of gasoline account for about 70% to 75% market share. DTW and Bulk sales of gasoline each account for about 10% to 15%. This market share distribution reflected in aggregated volume tables is that the publication cells for DTW and Bulk are much thinner (with fewer respondents) than those for Rack, and consequently the numbers of primary suppressions for DTW and Bulk are similar while both numbers are much higher than that for Rack. Table 1 above shows the numbers of primary suppressions defined by the 5% rule described above, and the total volumes suppressed by the primary suppressions.

To assess disclosure risk in the suppression patterns, the suppression patterns generated by the two software programs were audited using the Disclosure Auditing System (DAS) software developed by the Federal Committee of Statistical Methodology.[10] An audit of a suppression pattern produces upper and lower estimates for the value of each suppressed cell based on the data constraints. For this research project, any suppressed cells that could be estimated within 5% of their corresponding actual suppressed values were identified as disclosure cells. This means that the suppression pattern enabled a data user to estimate a suppressed value in the table within the target range of 5%.

Linear Programming Approach

Most software programs that automatically select complementary cells for suppression use a linear programming method that makes use of the hierarchical structure in the data.[11] Network flow methods may be characterized as a special case of linear programming. Routines based on network flow methods work well on two dimensional tables, with at most one level of hierarchy (in either rows or columns).[12]

When applying a linear programming methodology, sensitive cells are sorted and protected sequentially beginning with the most sensitive. For each sensitive cell, the set of complementary cells that minimize a cost function (commonly it is the sum of the suppressed values) is identified. Minimizing the sum of the suppressed values is one possible objective or cost function.[13] Other possible cost functions include minimizing

the total number of suppressed cells in a table or minimizing the suppression of specific data series in a table.

The software program used in this research is a modified software program used by the U.S. Census Bureau.[14] This program uses a minimal cost flow objective function to select complementary cells.[15] The program identifies specific closed paths, called "arcs" that protect a primary suppressed cell. Each line segment on the path in the network is assigned a cost per unit. For this research project, the cost is equal to the cell value. The capacity of a path is calculated for each arc. The capacity of each non-sensitive cell equals the required protection that the cell can provide to avoid closely estimating the value of a primary suppressed cell. For this study, the capacity is equal to the cell value. After each arc is assigned a cost and capacity, the program identifies a series of closed paths that give the least total cost and maximizes the capacity of the path to flow as may units as possible. The Minimal Cost Flow program uses an algorithm that reviews cell values to select the appropriate cell to suppress along a path to adequately protect a sensitive cell, moving in either a positive or negative direction along any closed path in a table. For this study, if a primary suppression has values R1 and R2 for its largest two respondents and if the other respondents in the cell have a total value of REM (the remainder of the cell), then the selected complementary cells must have a combined value that exceeds (R1) (5%) – REM + 1. The combined value of the complementary cells is the measure of the required protection for the primary suppression.[16]

Petroleum Marketing Programming Approach

Automated software was developed during the 1980s to perform primary and complementary suppression on petroleum data published in the *Petroleum Marking Monthly* publication. After the individual cells of a table are tested for sensitivity by applying primary disclosure rules, the cell values and their corresponding data characteristics are loaded into multi-dimensional arrays for complementary suppression analysis. Each characteristic of the hierarchical structure of the data becomes a dimension in the data structure.

The program initializes two four-dimensional arrays called "N" and "SW." The SW array holds the status of the suppression switches for each cell in the table. The N array holds the corresponding volumes from the initial pre-publication file with no suppressions for each table cell in the SW array. The dimensions of both the SW and N arrays are subscripted by the variables AREA, FORMULATION, SALES_TYPE, and GRADE.

The SW array maintains an extra element for each marginal within a dimension to serve as a flag to indicate whether complementary suppression is needed for a row or column total (marginal) within a dimension. The flags for a dimension have a value of "0" if no cells are suppressed in the dimensional array, a value of "1" if one cell is suppressed, and a value of "2" if two or more cells are suppressed. The flags function as signals for when complementary suppression is required in a specific dimension of the SW array. For example, if a flag has a value of "1," additional suppression is required for that dimension. Each dimension of the array has its own set of flags. The flags have a complex structure for the AREA, GRADE, and SALES_TYPE dimensions. For example, the AREA dimension contains nine separate flags because there are nine geographic marginals: The U.S. total, five regional totals, and three sub-regional totals. The SALES_TYPE dimension has three retail flags (total retail sales, sales through company operated outlets, and total sales to other end users) and four wholesale flags (total wholesale sales, Bulk sales, Dealer Tank Wagon sales, and Rack sales).

Once the data are loaded into both the N and SW arrays, complementary suppression is applied by reviewing each dimension of the SW array first. The program continues to select cells for complementary suppression until no marginal is left with only one suppressed cell. The program attempts to minimize the number of suppressed cells in a table by searching for the cell which has the highest number of intersections across the dimensions. The N array is reviewed to resolve ties between two cells which have the same number of intersections. An intersection occurs in the SW array when one cell may be selected as a complementary cell in two or more dimensions. Cells in the N array that are empty or zero are not included as candidates in the SW array. The complementary cells to be suppressed are selected, one at a time, by counting the number of unsuppressed cells in the each dimension. During each pass, the coordinates of the cell with the highest number of intersections are held by the program. The coordinates of the cell which is the best candidate as a complementary suppression are updated as the program identifies other cells with a higher number of intersections across the dimensions or a lower N value if the two or more cells have the same number of intersections. Once a cell is selected, the SW flags are cleared and the program continues to review the other dimensions within the SW array. The program reviews each dimension one at a time until the flags indicate that no additional suppression is needed along any dimension.[17]

Results from Applying Linear Programming Approach

The results from applying the linear programming approach and the PMM suppression program are shown in Tables 2 and 3. Table 2 shows that across each month, the total number of cells suppressed within each category remains fairly consistent: between the three months, there are 37-42 suppressed cells in the DTW sales category, the number of suppressions in the Rack sales category stays between 17 and 21 cells, and the number of suppressions in the Bulk sales category range from 23-26. This is important because it shows the stability of the data at the state level for these categories. If the initial selections of complementary suppressions are chosen to protect the cells that require the most protection, then the suppression pattern selected to protect the sensitive information in the table is generated within predictable outcomes for subsequent publication cycles. The stability of the suppression pattern over time is an important attribute for maintaining the data utility in a monthly publication.

It is interesting that while the number of cells suppressed is stable, there is a large amount of fluctuation in the volume of the suppressed data in each of the wholesale categories. The volumes in Table 2 are in thousand barrels per day. Table 2 shows that within the DTW category, approximately 22,000 volumetric units were suppressed in both July and October, while 40,839 volumetric units were suppressed in March, despite suppressing a lower number of non-sensitive cells. Additionally, in the Bulk category, there is a wide range in the total volume of suppressed units from 29,044 to 68,225 between March and October despite the fact that October has only one more suppressed cell. There is also a 4,000 unit difference between the data in March and July in the Rack category, despite having the same number of cells suppressed.

Table 2. Counts and Volumes of Complementary Suppressions

	DTW		Rack		Bulk	
	N	Volume	N	Volume	N	Volume
March						
DiAna	37	40839	17	15353	25	29044

PMM	56	117495	123	525993	28	71366
July						
DiAna	42	21607	17	11760	23	40894
PMM	65	114108	122	360558	34	120721
October						
DiAna	40	22243	21	24436	26	68225
PMM	63	118548	128	328903	28	101875

Results from Applying Petroleum Marketing Programming Approach

Table 2 shows that the complementary suppression patterns by DiAna contain fewer cells than those by the PMM suppression program. For Bulk the figures from the two systems are not that apart, in particular for the March and October data sets. For the DTW data, the PMM suppression programs flagged about 50% more cells that DiAna did. The Rack situation caught our eyes because the complementary suppressions from the PMM suppression program are much higher than those from DiAna. As described above, the linear program approach utilized in DiAna tries to minimize the total suppressed volumes while identifying potential complementary suppressions, and the PMM suppression program mainly focuses on minimizing the total number of suppressed cells. The total suppressed volumes of the PMM figures are all significantly higher than those DiAna figures.

Even though the criteria and approaches the two suppression systems use are different, many cells are selected by both systems as complementary suppressions. Except for the March Rack data, at least half of the complementary suppressions identified by DiAna are also part of the complementary suppression patterns from the PMM suppression program. In Table 3, the row stub labeled “both” refers to the condition where both programs selected the same cells for complementary suppression.

Combining the percentages of complementary suppressions selected by both systems and by the PMM suppression program alone, the PMM suppression program flagged at least 39% of the available cells as complementary suppressions. These figures along with the contrast shown in table 3 and 4 between the two suppression systems indicate the PMM suppression program might over flagged complementary suppressions, especially in the Rack data category.

In the research approach section it is mentioned that several years ago the marginal totals across sales type was removed from the publication tables in an effort to eliminate disclosure risk. But due to the complexity and inflexibility of the PMM suppression program, it is very hard to modify the program to run suppression on a data file without that marginal column. As a result, the whole data file including that marginal column was fed into the PMM suppression program during the selection stage of complementary suppressions. And then the whole column is removed from the publication. This practice certainly reduces disclosure risk of the released data, but it might flag some complementary suppressions unnecessarily in the reference to the suppression patterns with that marginal column not in the feeding data file. Without that marginal column the three categories of DTW, Rack, and Bulk are completely independent each other; and with that column in the feeding data file, the three categories are related and protection across sales type must be considered.

Table 3. Matches and Non-Matches of Complementary Suppressions			
	DTW	Rack	Bulk
	N	N	N
March			
Both	20	6	16
DiAna only	17	11	9
PMM only	36	117	12
July			
Both	29	14	17
DiAna only	13	3	6
PMM only	36	108	17
October			
Both	32	19	21
DiAna only	8	2	5
PMM only	31	109	7

After running both software programs on the three months of data, the results were analyzed by a data auditing program. The program audits the suppression pattern of a table by analyzing how well the suppressed cells are protected by estimating the actual cell values within a certain range. In this case, the program tested the results to see whether it could determine a range of values within 5% of the actual value because that was the parameter value set for applying the *p*-percent rule in this study. The results of auditing the suppression pattern generated by both the linear programming and petroleum marketing software are shown in Table 4.

Table 4. Disclosures Cells at 5% Protection Range						
	DTW		Rack		Bulk	
	DiAna	PMM	DiAna	PMM	DiAna	PMM
March	15	16	4	11	8	9
July	4	7	3	11	9	5
October	13	5	3	9	6	9

The auditing program generates and upper and lower bound estimate for each suppressed cell using all the data constraints among the published and suppressed values in the tables. In particular, the additive relationships among the non-marginal total cells and the marginal cells, and the non-negative nature of volume data. A suppressed cell is flagged as a disclosure if the estimated value for that cell is within the preset 5% protection range of the actual value. The auditing results for the three months show that the suppression patterns from the PMM suppression program contain more disclosure cells than that of the DiAna program for two out of three months for the DTW and Bulk sales categories and for all three months in the Rack sales category.

Conclusion

The suppression patterns generated by the two software programs were similar in their selection of complementary cells to suppress. The audit showed that the DiAna program

provided a better protection level over all three months. The PMM program suppressed more cells than the DiAna program. There was a programming mistake in the petroleum marketing suppression program that caused more suppression in the tables because the file structure indicated that totals were published for the wholesale sales category. The comparison between the software programs would been more complete if this correction was made to the program. The results show that the DiAna software suppressed fewer non-sensitive cells in applying a linear programming approach than the petroleum marketing suppression program.

---

[1] Robertson, Dale A., and Ethier, Richard. "Cell Suppression: Experience and Theory." p. 9.

[2] Id.

[3] Cox, Lawrence. "On Properties of Multi-Dimensional Statistical Tables."

[4] "Statistical Policy Working Paper 22 (Revised 2005)- Report on Statistical Disclosure Limitation Methodology, Federal Committee on Statistical Methodology.

[5] Id. p. 61.

[6] Id. p. 68.

[7] Tables 39 and 40, Petroleum Marketing Monthly publication available at http://www.eia.gov/oil_gas/petroleum/data_publications/petroleum_marketing_monthly/pmm.html

[8] The terms "Conventional" is defined as finished motor gasoline not included in the oxygenated or reformulated gasoline categories and excludes reformulated gasoline blendstock for oxygenate blending (RBOB. The term "Reformulated" is defined as motor gasoline formulated for use in motor vehicles, the composition and properties of which meet the requirements of the reformulated gasoline regulations promulgated by the U.S. Environmental Protection Agency under Section 211(k) of the Clean Air Act. The term "All Formulations" is the sum of both Conventional and Reformulated.

[9] The term "Dealer Tank Wagon" is defined as wholesale sales of gasoline priced on a delivered basis to a retail outlet. The term "Rack" is defined as wholesale truckload sales or smaller of gasoline where title transfers at a terminal.

[10] The Federal Committee on Statistical Methodology (FCSM) is an interagency committee that operates under the oversight of the Office of Management and Budget. It consists of 20 members from various federal agencies. The Committee investigates problems which affect the quality of Federal Statistical data, and makes recommendations and develops tools for improving statistical methodology in Federal agencies. http://www.fcsm.gov/about/ Date accessed 9/14/2011.

[11] Sande, Gordon. "Automated Cell Suppression to Preserve Confidentiality of Business Statistics. Stat. Jour U.N. ECE2 p. 33-41 (1984).

[12] "Statistical Policy Working Paper 22 (Revised 2005) - Report on Statistical Disclosure Limitation Methodology, Federal Committee on Statistical Methodology. p.68.

[13] Cox, Lawrence H. "Vulnerability of Complementary Cell Suppression to Intruder Attack."

[14] Zayatz, Laura. "Using Linear Programming Methodology for Disclosure Avoidance Purposes," Statistical Research Division Report Series, Census/SRD/RR-92/02 (1992).

[15] Jewett, Robert S. "Disclosure Analysis For Publication Tables," p. 11, (1992).

[16] Id. p. 5.

[17] Griffey, Michael J., Bournazian, Jacob, "Disclosure Avoidance Techniques Used in Petroleum Marketing Data." (1995).