Use of an Audit Program to Improve Confidentiality Protection of Tabular Data at the Bureau of Labor Statistics

Randall Powers and Stephen Cohen, Bureau of Labor Statistics

Keywords: Individually Identifiable Data, Audit, Confidentiality, Tabular Data

I. Introduction

The Bureau of Labor Statistics (BLS) is the main collector and provider of data for the Federal Government in the broad field of labor and economic statistics. BLS conducts a wide variety of establishment surveys to produce statistics on employment, unemployment, compensation, employee benefits, job safety, and prices for producers, consumers, and U.S. imports and exports. Data are collected from the full spectrum of establishments including manufacturers, retailers, services, state employment agencies, and U.S importers and exporters of goods and services. In an effort to prevent disclosure of individually identifiable data the tabular data are subjected to disclosure analysis algorithms which ensure that data users outside the Bureau can't get to individually responded data. The algorithms determine sensitive cells based on certain rules and suppress cells meeting those criteria prior to publication. Further, the algorithms identify necessary complimentary suppression cells to prevent derivation of primary suppressed cells via mathematical relationships in the tables.

Implementing cell suppression techniques optimally is an L-P hard computer application. Agencies have developed heuristic disclosure algorithms to determine and suppress confidential data. These procedures could contain deficiencies such that through complex mathematical means (e.g., linear programming methodologies), data users might be able to determine with great accuracy some of the suppressed cell values within the publications (Zayatz,1992). It can be shown that, using linear programming methodologies, an auditing system can be developed that evaluates the success of heuristic disclosure algorithms to protect individually identifiable data from disclosure.

In this paper we study the effectiveness of disclosure protection algorithm currently employed for the Quarterly Census of Employment and Wages (QCEW). The Office of Survey Methods Research (OSMR) conducted a disclosure audit of 2002 First Quarter data from for the state of Maryland. Analysis will be done using the Disclosure Audit System (DAS) software, a software application funded by six Federal Statistical Agencies including the Bureau of Labor Statistics.

II. Background

Currently, all data released by programs at the Bureau of Labor Statistics are subject to heuristic disclosure analysis algorithms which ensure that data users outside the Bureau can't ascertain the values of individually respondent data. The QCEW program publishes quarterly and annual counts of employment and wages reported by employers covering 98 percent of U.S. jobs, available at the national, state, Metropolitan Statistical Area (MSA), and county levels by North American Industry Classification System (NAICS) codes (BLS Handbook of Methods, 1997). An overview of current QCEW disclosure methodology follows.

Primary Nondisclosure

In primary nondisclosure, the estimation cells are evaluated to determine if releasing the data would enable a data user to estimate the value of an individual reporter too closely. If the values of an individual reporter can be estimated too closely, the cell is marked for suppression. Calculations based on microdata prepare the aggregated cells for primary nondisclosure. Cells are then evaluated based on the number of establishments in the cell, the amount of employment in the cell, the number of employers in the cell, and the contribution of the largest employers in the cell to total wages and employment. Subsequent runs of quarterly and annual files preserve the original nondisclosure flags.

A cell undergoes each of the following primary disclosure tests:

1. Employment dominance
2. Wage dominance
3. Establishment threshold
4. Employment threshold
5. Employer threshold

The p-percent test is used to determine cell sensitivity for both employment dominance and wage dominance (FCSM Statistical Working Paper 22, 1994). QCEW uses a version of the p-percent test which requires the data sums for the largest and second largest contributors to the cell, the data sum of the entire cell, and the p-value being tested against.

We can represent X_1 as the largest contributor, X_2 as the second largest contributor, and X as the entire cell. Using these three values, the difference ($X-X_1-X_2$) represents the residual data sum, that is, the sum of all but the largest two contributors to the cell. Under the p-percent test, the ratio of the residual sum to the data value X_1 is compared to the ratio represented by the p-value of the test. If the ratio $(X-X_1-X_2)/X_1$ is greater than or equal to the p-value, it is considered that the residual provides at least p-percent protection to the largest contributors value (X_1), and the cell data value X is considered discloseable.

If employment level of a cell or total wages is found to be sensitive, then all other data fields associated with the cell are suppressed as well. Total wages is checked for sensitivity when the employment cell is found not to be sensitive and will be suppressed if necessary.

In addition to the p-percent test, various threshold rules are also employed. These rules apply to number of establishments, number of employees, and number of employers.

On quarterly files, any cell with fewer than (small, unpublishable number) establishments for the quarter will be marked for suppression of the quarterly data. On annual average files, any cell with fewer than (slightly larger number) establishments will be similarly marked for suppression of the annual data.

On the quarterly files, any cell with fewer than (small, unpublishable number) employees for the third month of the quarter should be marked for suppression of the quarterly data. On the annual average file, any cell where the sum of all twelve months of employment is less than (slightly larger number) will be similarly marked for suppression of the annual data.

On the quarterly files, any cell with fewer than (small, unpublishable number) employers for the quarter will be marked for suppression of the quarterly data. On the annual average file, any cell with fewer than (small number) employers will be similarly marked for suppression of the annual data.

When a quarter is subsequently reprocessed after data have been released to the public, it is generally required that all cells suppressed in the prior release or releases be suppressed in the subsequent release.

Secondary Nondisclosure

For QCEW data, the cumulative data for cells at one level in a hierarchy are readily compared to the corresponding aggregate at the next higher step in the hierarchy. Secondary disclosure processing is necessary to prevent the ready determination of data for sensitive cells suppressed by arithmetic calculation using other cells in the hierarchy.

The QCEW disclosure suppression algorithm in use requires that there not be any situation in the data where, in the grouping of aggregate records at one level in a hierarchy and its component records at the next lower level in the hierarchy, there is only one suppressed record. In those situations, the secondary disclosure processing system is required to identify and mark at least one other cell for suppression. The rules for selecting such a cell are discussed further below.

The secondary nondisclosure system must simultaneously protect cells in all dimensions of processing. These include: the ownership-industry dimension, the area dimension, the size dimension, and the time dimension

There is a nine-level hierarchy of aggregation in the ownership/industry dimension. These levels are the following:

- Total (across ownerships)
- Total (by ownership)
- Geographic Region (by ownership)
- Alternate aggregate sector (by ownership)
- NAICS Sector (by ownership)
- 3-digit NAICS (by ownership)
- 4-digit NAICS (by ownership)
- 5-digit NAICS (by ownership)
- 6-digit NAICS (by ownership)

Note: Ownership may be by federal, state, or local governments, or by private industry.

For several of the area categories, there will be data cells at each of these ownership/industry hierarchy levels. For some of the categories, however, there are data cells at only some of ownership/industry hierarchies.

The area dimension is a more complex hierarchy than the ownership-industry dimension. At the lowest level are the county data, which aggregate to the statewide level in the hierarchy. The statewide data level in the hierarchy are aggregated (conditionally, in that they exclude Puerto Rico and Virgin Islands data) to the national level.

The county data, however, are also aggregated to the Metropolitan Statistical Area MSA/ Primary Metropolitan Statistical Area (PMSA)/ New England County Metropolitan Area (NECMA) level. Since the MSA/PMSA/NECMA aggregates sometimes span state borders, there are additional relationships among the area hierarchies that are checked to ensure that sensitive suppressed data cannot be readily solved.

The size dimension has only two hierarchies, total, and disaggregated. Similarly, the time dimension has only two hierarchies, annual and quarterly. However, there is a concern about comparing preliminary to revised data and the possibility that preliminary data could be used to reveal information about individual employers. An example of this would be a cell with three employers included in preliminary data which has a fourth small employer to it in the final data. The data user would thus easily be able to determine the exact values for data submitted by the fourth employer.

The preference rules for secondary disclosure processing are somewhat complex. However, the general rule is that the preference is to select for complimentary suppression the cell with smallest nonzero employment among those available. One conclusion from this preference is that it implies that in looking among a group of components (at one level of aggregation) and their immediate aggregate, if the group needs an additional suppression in that dimension, then a component record is preferred as a complimentary suppression over the immediate aggregate.

This preference rule is straightforward in the industry dimension and size dimension. Employment size level is the dominant rule. It is supplemented by other rules of consideration for the area and time dimensions.

III. Objective

Disclosure protection for Average Employment and Total Wages data will both be considered. The state of Maryland was chosen as a representative state for analysis. This paper will report on analysis that has been completed for four counties; one urban, one suburban, and two rural. Analyses of additional Maryland counties are possible.

IV. Disclosure Auditing Software

Disclosure Auditing System software (DAS) was developed in 2000 to share across the statistical agencies in the Federal Statistical system. DAS is auditing system using linear programming methodologies that checks that confidential data are provided utmost protection from disclosure. DAS uses SAS LP programming methodologies to flag to the user the range of values an outsider can determine a suppressed cell to be.

V. Analysis Plan

To use DAS, the user must first take published tabular data and convert it to comma-separated-value (CSV) input files using a package such as Excel. These files include record types which describe the dimension and hierarchies of the rows and columns of the table, record types which indicate protection range, and record types which contain individual cell values (Users Guide, 2001).

These files are then used as DAS input files, where the PROC LP Optimizer is used to determine the largest and smallest values for a suppressed cell, given the cells and marginals that have been released. Software should estimate the narrowest gap between these two values that is possible given the table structure. This is desirable because you are thereby determining the tightest range in which the true value could fall. A narrow gap would more easily allow an outsider to guess or estimate the true value of the cell, hence, potentially determine the value of an individual reporter. If the program produces a range which must be wide, you are less likely to estimate a cell with a narrow gap.

The user specifies a protection range criteria for suppressed cells; that is the percentage above and below the actual suppressed cell value for which protection is desired. For this exercise, we chose 2.5 percent as the protection range.

A cell is deemed a problem cell if the gap between the largest and smallest possible value of the cell is smaller than the five percent protection range gap.

The Objective Functions

The purpose of the LP model is to solve what are known as objective functions, for maximums or minimums subject to constraints. For this auditing software, we do both. That is, we seek to determine both a maximum cell estimate and a minimum cell estimate around a tabular cell that has previously been suppressed using disclosure software conforming to certain rules.

Since it is our objective to determine estimates for suppressed tabular cells, the auditing software's first procedure after the input and verification methodologies is to set up all of the objective functions that the auditing software must solve given a table or tables of data. The software identifies all of the suppressed cells contained in the table. An objective function record for each suppressed cell is written to the SASDL (or specified output library).

The Constraints

Meaningful solutions to the objective function must satisfy a set of constraints, or a system of equations and constants which bound the limits of all values within a table. The Auditing System software also generates from the CSV imported data, all of the constraints that bound the solutions set, or objectives. Constraints consist of any unsuppressed cell in the table(s), including the margins (totals, and subtotals).

Optimization

Upon completion of generation of the table driven (data driven based on the values imported from the CSV file) objective functions and constraints, the LP procedure begins the LP optimization stage. The LP procedure provides the first (MAX) objective function (first suppressed cell) and all of the constraints to the LP optimizer. The optimizer then generates a base tableau (a set of complete base values for all cells, including suppressed cells) and writes these data to the SASDL library. This data set is used in subsequent optimizations of the remaining objective functions. (Note: the procedure conducts the same approach for the MIN objective functions). After generating the base tableau, the LP optimizer determines the optimized value for the objective function. The optimized value is then written to the LPMAX or LPMIN data sets in the SASDL library.

After solving the LP for the first objective function, the LP procedure runs the second objective function, constraints and the base tableau to the optimizer for processing. This procedure is performed as many times as there are objective functions. When all objective functions have been processed through the optimizer, the values also are written to a SAS data set named FINAL in the SASDL library.

VI. Example

For confidentiality reasons, we do not use actual BLS data for the example.

Therefore, all of the numbers in the following example are completely fictitious. The following table (see attachment 1) is used as input for the sample. (Note: Due to space issues, only part of the table is shown here.)

The first column represents the NAICS code of the cell. The second column is the published total wage figure for the associated cell, and the third column is the actual value. When a cell has been suppressed, either a P for Primary suppression or C for Complementary suppression is indicated in the second column.

CSV input files are created for each two-digit NAICS code. The files contain data for six, five, four, three and two digit NAICS codes, all of which sum to the next highest level of data (e.g. six digit codes sum to five digit codes, etc.). Two separate files are created for each two-digit NAICS code: one file for Total Wage data, the other for Average Employment data. Average Employment data is taken to be the third month of the quarter, rather than the three month average. Each file is then imported into the DAS Program, where the PROC LP Optimizer is used to minimize and maximize the linear function subject to linear constraints. For the purpose of this analysis, the protection range for the cells was specified as being within 2.5% above or below the actual cell value.

The DAS program then takes the input file and produces a file of objectives and constraints (Attachment 2, also only partially reproduced here). This attachment contains the following information:

Objectives:

OBJ13: NAICS 2331
OBJ16: NAICS 2339
Etc.

These are the objectives (i.e. the suppressed cells) that the DAS program is attempting to find a minimum and a maximum for.

Constraints:
CON1: 772.5 <= NAICS 23<= 773.5
CON9: 592.5 <= NAICS 233<= 593.5
.
.
CON30: NAICS 233–NAICS 2331-NAICS 23312 = 0
CON31: NAICS 2331–NAICS 23311 – NAICS 23312 = 0
.
.
These are the constraints, the set of equations which bound the objectives. There are two types of constraints. The first, exemplified by CON9 and CON10, simply take the cell values that are published (those not suppressed), and put rounding boundaries on them, for software operational purposes. The second type, exemplified by CON30 and CON31, express the row and column additivity constraints.

The Proc LP Optimizer is then used to minimize and maximize the linear function subject to these objectives and linear constraints.

The DAS program produces a final output dataset (Attachment 3). This contains the following information:

NAICS: North American Industry Classification System (NAICS) code.
Actual: Actual cell value that has been suppressed in the official publication
LB: Lower Bound (The value 2.5% below the actual cell value)
UB: Upper Bound (The value 2.5% above the actual cell value)
Min: Minimum value as determined by optimizer. This is the best estimate that can be determined of the minimum value of the cell.
Max: Maximum value as determined by optimizer. This is the best estimate that can be determined of the maximum value of the cell.
FIF: Notation which indicates if any minimized or maximized cells have been found. A cell is said to be minimized if the minimum value of the suppressed cell can be determined as a value greater than the lower bound (LB). Similarly, a cell is said to be maximized if the maximum value of the suppressed cell can be determined as a value less than the upper bound (UB). Findings of “minimized” or “maximized” are only problematic if the Feasibility Interval (Max-Min) is less than the Protection Range (UB-LB).
SF: Notation which indicates when the Feasibility Interval (Max-Min) is less than the

Protection Range (UB-LB). When this occurs, we have a disclosure violation.

The min and max are produced for each of the suppressed cells. The min is the minimum value that the cell could possibly be, and the max is the maximum value that the cell could possibly be, given the published cell structure and marginal totals. Software should estimate the narrowest gap between min and max that is possible given the table structure. This is desirable because you are thereby determining the tightest range in which the true value could fall. A narrow gap would more easily allow an outsider to estimate the true value of the cell. If the program produces a range which must be wide, you are less likely to estimate a cell with a narrow gap.

How are the minimum and maximum determined? Each suppressed NAICS cell becomes an objective function (see attachment 2) for which the LP Optimizer will try to solve for. The relationships between the NAICS codes determine linear constraints that the optimizer uses to attempt to solve for the objective function.

For example, suppose we wish to solve for OBJ13, which is NAICS code 2331. The constraints that affect this cell are:

CON9: 67.5 <= NAICS 233<= 68.5
CON10: 45.5<= NAICS 23312<= 46.5
CON30:NAICS 233–NAICS 2331 –NAICS 2339 = 0
CON31: NAICS 2331 –NAICS 23311 –NAICS 23312 = 0

Constraints 9 and 10 essentially tell us the following actual cell values:

NAICS 233=68
NAICS 23312=46

We can also replace known values in the following constraints:

CON30:NAICS 233–NAICS 2331 –NAICS 2339 = 0

To produce: a.) NAICS 2331+NAICS 2339=68

Also:

CON31: NAICS 2331 –NAICS 23311 –NAICS 23312 = 0

Produces: b.) NAICS 2331-NAICS 23311=46

From a.), we can see that cell 2331 is at most 68. This is our max.
From b.), we can see that cell 2331 is at least 46. This is our min.
The outside data user can, at best, guess that:

46<=NAICS 2331 <=68

The actual value for cell NAICS 2331 is 61.

To find the upper bound, we take (.025*cell value)+cell value. The upper bound is 62.525.

To find the lower bound, we take cell value-(.025*cell value). The lower bound is 59.475.

A cell is deemed a "problem cell" if (max-min)<(upper bound-lower bound). That is, if the outside user can predict the gap of possible values of the cell to be smaller than the gap as defined by the DAS user. In this example, this is not the case. Thus, the cell is found to have adequate protection.

Interestingly, even if the max were less than the upper protection bound, we would not have had a problem cell. This is because the outside user can only produce the gap between max and min, he could not possibly be aware of the potential closeness of the max to the actual value.

Attachment 3 shows the results file for NAICS code 23. For this particular input file, we see that in all cases, the gap between the min and max as determined by DAS software is larger than the gap between the lower and upper bounds. Thus, we have no problem cells.

VII. Analysis

A cell is said to be "maximized" if the max produced by the DAS software falls within the requested protection range. Similarly, a cell is said to be "minimized" if the min determined by DAS is within the requested protection range. A total of 4540 suppressed cells were analyzed. Of these, 185 (4%) were either maximized or minimized. A cell which is either maximized or minimized is not necessarily a problem cell.

The real problem arises when the feasibility interval (essentially the difference between the max and min as produced by the DAS software) for a primary cell is smaller than the requested protection range. This problem did not occur in any of the counties examined.

There were seven cells that failed this test, but they were all complementary cells. Since this will not affect the data users ability to predict primary cells, we are not concerned with those seven cell failures.

Additional analysis was performed at the 5%, 10%, and 20% levels. Problems with primary cells were not detected until the 20% level. Because the 20% level affords us 40% protection of the cells, a data user would not be able to come nearly close enough to obtaining the true values of suppressed cells.

VIII. Conclusions

In conducting this audit, we wanted to apply auditing software to validate suppression patterns as applied by the QCEW program.

The Disclosure Auditing System software is not designed to easily evaluate a survey that publishes tables for six digit deep NAICS codes for six million establishments in all fifty states. Thus, a representative sample was chosen to evaluate the adequacy of the suppression patterns for the QCEW.

Total Wages and Average Employment data for all NAICS codes of four Maryland counties were analyzed at the 2.5% level. Analysis showed that no problems were encountered for the four counties that were studied. Further analysis showed that no problems occurred until the 20% level, and this would not allow a data user to come nearly close enough to obtaining the true values of suppressed cells.

As resources and time permits, we will attempt to evaluate as many counties as possible.

IX. Acknowledgement

The authors would like to thank Michael Buso of the Office of Employment and Unemployment Statistics at BLS for providing background information on nondisclosure techniques as applied to the QCEW, as well as the QCEW data set used for analysis.

X. References/Bibliography

"CEW Nondisclosure System Requirements For NAICS." *BLS Internal Report*, January 7, 2002.

Federal Committee on Statistical Methodology. (2001)."*Federal Committee on Statistical Methodology Disclosure Auditing System Users Guide.*"

Federal Committee on Statistical Methodology (1994). "*Statistical Working Paper 22, Report on Statistical Disclosure Limitation Methodology.*" Washington D.C.: U.S Office of Management and Budget.

U.S. Department of Labor, Bureau of Labor Statistics. (1997). *BLS Handbook of Methods*.

Zayatz, Laura, (1992). "*Using Linear Programming Methodology for Disclosure Avoidance Purposes*." Bureau of the Census Statistical Research Division Research Report Series,RR92-02.

Attachment 1: Input File (partial)

NAICS Code	Published Total Wage	Actual Total Wage
23	773	773
233	68	68
2331	C	61
2339	P	7
23311	P	15
23312	46	46
23392	P	4
23393	P	3
232110	P	15
233120	46	46
233920	P	4
233930	P	3

Attachment 2: Objectives and Constraints (partial)

Objectives:
OBJ1: NAICS 23211
OBJ2: NAICS 232110
.
OBJ13: NAICS 2331
.
OBJ16: NAICS 2339
.
OBJ26: NAICS 234292

Constraints:
CON1: 772.5 <= NAICS 23<= 773.5
CON2: 592.5 <= NAICS 232 <= 593.5
.
.
CON9: 67.5 <= NAICS 233 <= 68.5
CON10: 45.5 <= NAICS 23312<= 46.5
.
.
CON30: NAICS 233-NAICS 2331-NAICS 2339 = 0
CON31: NAICS 2331-NAICS 23311-NAICS 23312 = 0
.
.
CON43: NAICS 23429-NAICS 234291-NAICS 234292 = 0

Attachment 3: Output File (complete)

NAICS	actual	lb	ub	min	max	FIF	SF
23211	165	160.875	169.125	0	182		
232110	165	160.875	169.125	0	182		
23212	16	15.6	16.4	0	182		
232120	16	15.6	16.4	0	182		
23221	88	85.8	90.2	0	99		
232210	88	85.8	90.2	0	99		
23229	20	19.5	20.5	9.5	108.5		
232292	10	9.75	10.25	0	99		
23231	14	13.65	14.35	0	23.5		
232310	14	13.65	14.35	0	23.5		
23239	9	8.775	9.225	0	23.5		
232390	9	8.775	9.225	0	23.5		
2331	44	42.9	45.1	34.5	49.5		
23311	9	8.775	9.225	0	15		
233110	9	8.775	9.225	0	15		
2339	5	4.875	5.125	0	15		
23392	4	3.9	4.1	0	15		
233920	4	3.9	4.1	0	15		
23393	1	0.975	1.025	0	15		
233930	1	0.975	1.025	0	15		
23411	13	12.675	13.325	0	27.5		
234113	13	12.675	13.325	0	27.5		
23412	14	13.65	14.35	0	27.5		
234126	14	13.65	14.35	0	27.5		
234291	3	2.925	3.075	0	6.5		
234292	3	2.925	3.075	0	6.5		