Evaluation of Proposed County-Level Interval Publication of ES-202 Employment Data

Office of Survey Methods Research

April 23, 2004

Abstract

According to the proposed changes to ES-202 publication policy for aggregate employment data where suppressed cells are published in pre-determined ranges, we found replacing primary and complementary suppression cells selected under current methodology with pre-determined ranges could narrow the intruder estimation of individual establishment employment levels in the cells that may pose confidentiality exposure risks. Current theoretical knowledge and techniques for selecting complementary suppression cells do not provide a direct solution. Neither any solution such that the information loss due to range publication is minimized while satisfying confidentiality protection rules exists. In this research, we evaluate a heuristically feasible method using currently available complementary suppression cell selection software Disclosure Analysis (DiAna) and list our findings by applying the heuristic to a subset of a ES-202 data set.

Interval Publication of Cells with Disclosure Risk

The Problems with Interval Publication

Under the proposed changes to ES-202 publication policy for aggregate employment data where suppressed cells are to be published in pre-determined intervals, similar to employment summaries published in Census Bureau's County Business Pattern, we found we may disclose more information about individual respondents in the primary suppression cells. The reasons are:

1. Publishing a range instead of completely suppressing the cell value provides additional information about the actual value of the cell to the intruder. Previously, the only mechanism a intruder utilizes is the column and row marginal in pursuing the estimates of other establishment actual employment values. Now, with published upper and lower bounds of supposedly completely suppressed cells, the intruder may obtain a closer estimate of the primary cells by taking into consideration of these published upper and lower bounds by comparing results from subtraction from column and row marginal obtained. This violated the p-percent rule, which states that none of the respondent values should be estimated precisely enough to within p-percent of its actual value by anyone, from outside or within the cell. Some examples are worked out to demonstrate this problem, see Li [2] and Ernst [1].

2. In certain cases where selection of complementary cells is irrelevant, an internal intruder within the primary suppression cell could narrow its estimation of other establishment employment within a closer range of its actual value than that is required by two-sided p-percent protection rule. This is independent of how complementary cells are selected. Internal intruder in the primary cell is able to narrow its estimates by just looking at the published boundaries, a range replacement of complete suppression reveals itself too much about the actual cell value. Further more, it is theoretically possible that the overall estimated interval by the intruder is narrower than what is required even though the

required protection of the primary cell is sufficiently provided through complementary suppression cells, see Li [3] for a description of this situation.

Current theory and techniques for selecting complementary suppression cells do not provide a direct solution such that the information loss due to proposed interval publication is minimized while also satisfies p-percent rule. The currently available automation tools for selecting complementary cells are all designed under the assumption of complete suppression. The source code reflects this method can not be modified directly to produce even an approximate solution for interval publication of suppressed cells, this is pointed out by one of the software authors[1]. Whether additional reasonable amount of work in this effort will succeed is unclear. If we want to check manually, there are in the magnitude of 2^n possible patterns of cell suppressions in a publication table, n is the number of cells in the table, each of the pattern needs to be checked whether required protection of primary cells are satisfied. The amount of work escalates quite fast with n. It becomes prohibitive to nether evaluating disclosure risk nor producing BLS publications.

Proposed Solution

In the method of this study, we propose that each of the cell capacity to protect primary suppression cells needs to be modified before feeding into DiAna, the software we could use to select complementary suppression cells under the complete suppression rule. The new capacity of a cell is defined to be the minimum of the absolute value of the difference between the actual value and the lower and upper bounds of the publication range.

For example, in the ijth employment publication cell of ES-202, let R1 be the largest respondent, R2 the second largest and Remainder the sum of the rest. We transform the cell values/capacities according to the new capacity b_{ij}:

$$b_{ij} = \min\{a_{ij} - l_{ij}, u_{ij} - a_{ij}\}$$

where

$\ell_{ij} =$ lower bound of publication interval for cell ij

$u_{ij} =$ upper bound of publication interval for cell ij

$a_{ij} =$ actual cell value for ij

This new capacity is used instead for the aggregated cells ranked lowest in the hierarchical tree of the industries by NAICS digits. Capacities of higher order cells are sums from the capacities of lower order cells. A new publication table completed with re-defined capacities is used for complementary cell selection in the existing software. The selected cells will be published in pre-defined intervals as those shown in the Appendix. The cells were not selected as complementary cells will be published with their actual values. Primary suppression cells are published also in intervals. Note here is another problem where method of selecting complementary can not solve. See Li [3] for more discussion on treatment of primary cells.

To use the above method is because current theoretical knowledge and techniques selecting complementary cell suppressions do not provide a solution such that the information loss (however it is defined) is minimized with range publication of suppressed cells while protecting confidentiality. Therefore we need to modify existing method to select range publication complementary cells. The original method in the software uses network flow and minimum cost flow (MCF) techniques. This method for 1-d or 2-d tables that is expected to produce an interval publication that will satisfy the two-sided p-percent confidentiality protection requirement while being close to an optimal solution. The modified is one among possible solutions to the interval publication problem, this method is relatively less complex to implement with existing automation tool. In the following section we describe an empirical study of approximate loss of

[1] Dandekar, Ramesh A., Energy Information Administration, Ramesh.Dandekar@EIA.DOE.GOV

information between complete suppression and interval publication of complementary cells using ES-202 Maryland employment data. For detailed justifications of this method see Ernst [2] and Li [1].

Intended Results:

1. All primary cells are protected (in the sense of ``p-percent rule''),
2. The number of cells published in intervals under the new proposal will be more than the number of cells completed suppressed for disclosure avoidance.
3. This is not the optimal algorithm to select cells for fixed interval publication (in the sense of minimizing ``information loss'').

The Procedure to Compare Two Disclosure Avoidance Methods

We use the State of Maryland first quarter 2002 establishment data to evaluate these two selection methods. This file contains 35,720 establishments with a total employment level of 2,411,755 in twenty four counties (including Baltimore City) in the State of Maryland. We select eight 2-digit NAICS supper sectors to limit the amount of data so that we could speed up the evaluation process. The eight supper sectors are:

2-digit NAICS	Industry	Total Employment
31-32	Manufacturing	54,244
44-54	Retail Trade	332,011
48-49	Transportation and Warehousing	65,224
51	Information	37,650
52	Finance and Insurance	137,511
53	Real Estate and Rental and Leasing	80,104
54	Professional, Scientific and Technical Services	289,659
62	Health Care and Social Assistance	169,985
	Total	1,166,388

The total employment of these eight super sectors accounts for about half (48.4%) of total Maryland employment. These eight supper sectors cover about half (46.2%) of all establishments in Maryland:

2-digit NAICS	Industry	Total Number of Establishments
31-32	Manufacturing	1,008
44-54	Retail Trade	4,356
48-49	Transportation and Warehousing	864
51	Information	451
52	Finance and Insurance	1,764
53	Real Estate and Rental and Leasing	1,078
54	Professional, Scientific and Technical Services	3,871
62	Health Care and Social Assistance	3,151
	Total	16,527

The computing ability to process these micro-level data from the entire State of Maryland at once is limited, even with this partial selection of NAICS industry groups. To process the data we first need to produce the publication table in a structure similar to the official ES-202 release. Secondly we produce a parallel publication table with the largest and second largest establishments at NAICS 3 to 6-digit levels by county. This is for one of the four input files of DiAna, the software utilizes MCF methodology to select suppressions cells. Thirdly, hierarchical row relationships of the publication tables has to be identified and clearly arranged in a flat file for DiAna input. We also need to re-assign the "capacity" of publication cells in order to evaluate the proposed Interval Publication method. S-plus scripts are written to process data in above steps. The total amount of data to be processed exceeds the top amount of PC memory allocated for the software. This is why we have to further dissect the 2-digit NAICS industry codes into 3-digit NAICS codes and analyze at the 3-digit level.

The eight 2-digit industry supper sectors, are analyzed at and below their 3-digit NAICS industry levels. Maryland has 61 3-digit NAICS industries. The data is first grouped under these 61 divisions. For each of the division, we proceed from steps 1 through 3 described in the previous paragraph. Once the input files are ready, we use Diana to select the suppression cells at the final step.

We assume the Maryland employment data are published in NAICS hierarchical order from 6-digit up to 3-digit for each county and the State of Maryland total, which is a marginal sum of employment levels of

every county in Maryland. The publication table is therefore 2-D with one dimension (NAICS) in hierarchical order.

For the Complete Suppression method, which completely suppresses a cell if the cell is identified as breaking confidentiality rule, the p-percent rule, a routine has been developed and implemented in DiAna. For the proposed Interval Publication, which requires cells to be published in pre-determined intervals while protecting the confidentiality of individual establishments, we developed a method to select cells that is indirectly implemented in DiAna. This method recalculates the “capacity” of the publication cells before they feed to the DiAna input. For details see references at the end of this note.

S-plus is a general statistical software produced by Insightful Corporation. Statisticians mainly use it to do statistical modeling and data analysis. However its scripting language is also very powerful for data processing. Here we use it to finish majority of data processing job.

This evaluation is done at and below 3-digit NAICS level. In theory it is possible that some 3-digit NAICS cells at county-level have to be suppressed, though in practice may be rare. This is one of the drawbacks from dividing the whole state data along 3-digit industry lines that we have to because of limited computing capacity.

Comparison of Suppression Patterns under Two Methods

Here is what we found after evaluating the suppression patterns under the current Complete Suppression method and the proposed Interval Publication method:

1. *Interval Publication method selects about three times more either the level of employment or the number of cells to be published in pre-defined intervals than what Complete Suppression chooses to completely suppress*. In the State of Maryland, this is 34.4% of total employment vs. 13.3%, and 36,4% of publication cells vs. 11.9%. This is consistent with the five-county study we did earlier at the 2-digit NAICS industry level.

2. *Variation of percent of employment and publication cells suppressed across 3-digit NAICS industries is small*, given the establishment employment patterns in different industries could be quite different, and the methods used to select the suppression cells are different. This in part suggests the methodological consistency underlying the two methods used to select suppression cells.

3. *The difference between percent employment and percent cells suppressed is small*. There is no clear pattern which is consistently larger in either of the two methods, though it seems for larger 3-digit NAICS industry the percent of cells suppressed tend to be larger. This may be caused by the heterogeneous small and large business mix in these industries that forces cells with dominate companies to be suppressed. Larger industry groups also cover more detailed sub-industries and wider geographical coverage that may lead to sparse cells that prone to suppression.

4. *The overall information loss between the two suppression methods is not compared*. Exactly how much information is gained by publishing a cell in pre-determined cells over completely suppress it is not defined yet. This is difficult because the cells suppressed under one method are not those selected to be published in intervals under another method. We need an overall measure of information delivered through a publication table in order to measure the difference of information loss between two methods.

The current study is suggestive to how the publication table would look like. However it does not measure which method provides more information to the reader within constrains of the confidentiality requirement. Even though Interval Publication may give more cells in vague terms, it could well be that by publishing the cells in intervals some readers see greater usefulness of this portion of the BLS data.

Detailed outputs are included in Table 1,2,3, and Figure 1, 2

Table 1. Suppression Pattern Under Complete Suppression Method (Maryland)

	Publication Employment			Publication Cell		
NAICS Code	Suppressed	Total	%	Suppressed	Total	%
311	587	4513	13	105	1166	9
312	40	331	12	34	286	12
313	138	1149	12	39	357	11
314	301	2148	14	37	286	13
315	66	553	12	41	405	10
316	11	70	16	13	95	14
321	89	591	15	53	381	14
322	266	2214	12	37	333	11
323	858	9530	9	46	357	13
324	18	108	17	17	167	10
325	315	2864	11	97	881	11
326	198	1802	11	53	405	13
327	198	1801	11	68	619	11
331	116	724	16	71	643	11
332	784	6533	12	168	1119	15
333	436	4839	9	120	1095	11
334	293	3252	9	97	809	12
335	48	479	10	54	452	12
336	159	1322	12	120	857	14
337	343	3425	10	62	476	13
339	540	5996	9	57	571	10
441	3468	38530	9	60	428	14
442	1965	16373	12	29	238	12
443	2123	23590	9	26	214	12
444	2639	23987	11	46	357	13
445	4895	54394	9	67	476	14
446	2125	26568	8	29	262	11
447	1827	22837	8	16	143	11
448	4068	36983	11	47	524	9
451	2398	18449	13	42	381	11
452	1482	13474	11	33	238	14
453	4908	49081	10	79	524	15
454	1162	7745	15	50	381	13
481	173	1330	13	24	238	10
483	74	460	16	21	190	11
484	3101	23855	13	37	333	11
485	1050	9541	11	54	452	12
487	138	1252	11	25	167	15
488	2390	19917	12	93	666	14
491	4	34	11	10	95	11
492	460	4180	11	18	167	11
493	605	4655	13	21	190	11
511	829	10358	8	50	357	14
512	514	5138	10	54	357	15

515	155	1108	14	24	238	10
516	14	139	10	12	95	13
517	1716	12256	14	57	405	14
518	971	8089	12	19	190	10
519	90	562	16	29	190	15
522	7120	47467	15	73	666	11
523	2243	16019	14	77	547	14
524	10147	72476	14	61	405	15
525	217	1549	14	41	405	10
531	8769	58459	15	45	500	9
532	1922	21361	9	104	690	15
533	20	284	7	11	95	12
541	49242	289659	17	214	2142	10
621	17414	124384	14	102	1023	10
622	102	681	15	26	238	11
623	2394	17101	14	34	381	9
624	3895	27819	14	73	524	14
Total	*154633*	*1166388*	*13.3*	*3322*	*27872*	*11.9*

Table 2. Replacement Pattern under Partial Interval Suppression Method (Maryland)

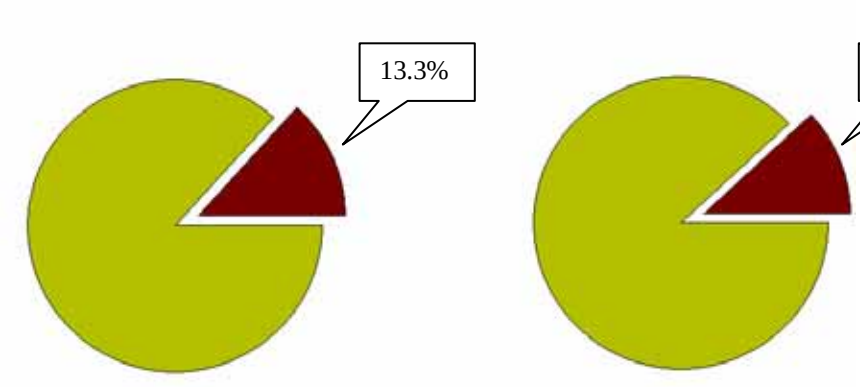
	Publication Employment			Publication Cell		
NAICS Code	Suppressed	Total	%	Suppressed	Total	%
311	1670	4513	37	420	1166	36
312	106	331	32	109	286	38
313	379	1149	33	129	357	36
314	730	2148	34	112	286	39
315	177	553	32	150	405	37
316	22	70	32	35	95	37
321	230	591	39	141	381	37
322	775	2214	35	117	333	35
323	3240	9530	34	129	357	36
324	39	108	36	58	167	35
325	1088	2864	38	308	881	35
326	613	1802	34	142	405	35
327	576	1801	32	204	619	33
331	282	724	39	244	643	38
332	2287	6533	35	369	1119	33
333	1790	4839	37	416	1095	38
334	1041	3252	32	291	809	36
335	182	479	38	154	452	34
336	529	1322	40	326	857	38
337	1267	3425	37	186	476	39
339	1919	5996	32	194	571	34
441	12330	38530	32	154	428	36
442	5731	16373	35	81	238	34
443	9672	23590	41	79	214	37
444	8156	23987	34	136	357	38
445	17950	54394	33	186	476	39
446	9564	26568	36	97	262	37
447	8678	22837	38	47	143	33
448	12574	36983	34	189	524	36
451	7195	18449	39	126	381	33
452	4312	13474	32	88	238	37
453	16197	49081	33	204	524	39
454	2478	7745	32	141	381	37
481	479	1330	36	83	238	35
483	147	460	32	65	190	34
484	7634	23855	32	127	333	38
485	3149	9541	33	172	452	38
487	388	1252	31	65	167	39
488	6373	19917	32	240	666	36
491	12	34	34	33	95	35
492	1714	4180	41	55	167	33
493	1676	4655	36	63	190	33

511	4040	10358	39	129	357	36
512	1952	5138	38	136	357	38
515	355	1108	32	79	238	33
516	53	139	38	35	95	37
517	4902	12256	40	146	405	36
518	2831	8089	35	65	190	34
519	185	562	33	68	190	36
522	17088	47467	36	233	666	35
523	6247	16019	39	197	547	36
524	23192	72476	32	146	405	36
525	604	1549	39	146	405	36
531	22214	58459	38	190	500	38
532	6622	21361	31	269	690	39
533	111	284	39	35	95	37
541	89794	289659	31	835	2142	39
621	49754	124384	40	348	1023	34
622	245	681	36	90	238	38
623	6498	17101	38	137	381	36
624	8624	27819	31	199	524	38
Total	*400662*	*1166388*	*34.4*	*10148*	*27872*	*36.4*

Table 3. Overall Comparison of Two Methods

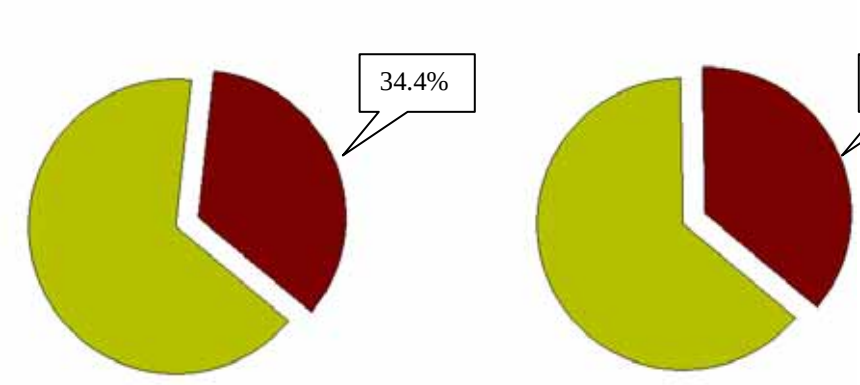
	Employment Suppressed/Replaced		Cells Suppressed/Replaced	
Method	Number	Percentage (%)	Number	Percentage (%)
Complete suppression	154633	13.3	3322	11.9
Interval publication	400662	34.4	10148	36.4

Figure 1. Method of Complete Suppression



Published Total Employment Suppressed Published Summary Cells Suppressed

Figure 2. Method of Interval Publication



Total Employment Published in Intervals Summary Cells Published in Intervals

Appendix

Publication interval used by Census Bureau:

A: 0-19
B: 20-99
C: 100-249
E: 250-499
F: 500-999
G: 1000-2499
H: 2500-4999
I: 5000-9999
J: 10,000 to 24,999
K: 25,000 to 49,999
L: 50,000 to 99,999
M: 100,000 or more.

Reference:

[1] Ernst, L.(11/13/03) *Comments on "A Method for Selecting Complementary Cell Suppressions for Fixed Interval Publication and Nondisclosure Estimates"*. Unpublished internal manuscript, BLS.
[2] Li, B.T. (9/03) *A Method for Selecting Complementary Cell Suppressions for Fixed Interval Publication and Nondisclosure Estimates*. Unpublished internal manuscript, BLS.
[3] Li, B.T. (11/18/03) *Situation Where sliding Interval Rule is Necessary for Primary Cell Protection*. Unpublished internal manuscript, BLS.