Comments on Cell Suppression Methods and Comparisons of Cell Suppression Software

*Gordon Sande, Sande and Associates, Inc.**

The protection of respondent confidentiality by withholding selected cells in magnitude tabulations of official economic statistics is standard practice. This practice has been automated by several statistical agencies. The mathematical framework for this practice is straightforward, although care must be taken to ensure that the details are carried through consistently. Some recent papers have problems with their details. The software to implement cell suppression is also straightforward, although it is highly technical and draws methods from areas unfamiliar to statisticians. A comparison of three systems for cell suppression has been carried out on the same set of live economic microdata. The comparison involves both the ability of the systems to protect the respondents' confidentiality and to preserve the utility of the economic tabulations.

■ Introduction

The protection of respondent confidentiality by withholding selected cells in magnitude tabulations of official economic statistics is standard practice. This practice has been automated by several statistical agencies. There have been several suggestions that there should be a comparison of the automated software from the U.S. Bureau of the Census (USBC) and from Statistics Canada. In carrying out and reporting on such a comparison, a number of areas of concern arise. Some of the concerns are specialized and highly technical. Other concerns are general and related to the basic formulations and to the properties of the data. In this note, we will look at several of these basic concerns.

■ Data

The standard example problem for cell suppression is a tabulation from a census of manufacturing. The tabulation reports the value of manufacturing shipments for varying geographical areas and for varying Standard Industrial Classification (SIC) codes. The microdata for this tabulation have four fields. Schematically, they are:

SIC	Geo	CoId	Data

SIC--Standard Industrial Classification
- type of business activity
- advertised by business

Geo--Geography
- location of business activity
- advertised by business

CoId--Company Identification
- ownership of business activity
- advertised by business

Data
- amount of business activity
- confidential to business

The *CoId* field would not be present in data from a similar census or survey of persons. This identification field for common ownership is a characteristic of business data. It allows the separate establishments to be grouped together to form enterprises. Cell suppression is based on a notion of sensitivity or concentration that is calculated from the ordered sizes of the enterprises that contribute to a cell. The practical difficulty that arises is that, when we aggregate cells, the identity of the largest contributing enterprise may change. The value of the total is additive under aggregation, but the count of enterprises is subadditive. In this simple example, we aggregate two cells of three enterprises to get a single cell of five enterprises. Sue is the largest contributor in the aggregate cell but not in the separate cells.

10	50	60	120
Tom	Sue	Dick	Total

+

5	45	75	125
Harry	Sue	Joe	Total

=

5	10	60	75	95	245
Harry	Tom	Dick	Joe	Sue	Total

■ Sensitivity Measures

A common traditional sensitivity measure of a cell is its concentration. If we were to say that any cell in which the largest two contributors form more than 75 percent of the total is sensitive, then we would have the form

$$S(x) = 1/3 \{ x(1) + x(2) \} - x(3+)$$

as our sensitivity function. We use x(1) for the largest, x(2) for the second largest, and x(3+) for the total of the third and higher largest contributors. If we seek a sensitivity measure that limits the ability of one of the respondents to estimate the values of the other respondents, then we would have the form

$$S(x) = 1/3\ x(1) + 0 - x(3+)$$

where we limit the improvement to three times better than it would be without the publication of the cell total. Generalization for other parameter values for both forms can be readily done. The concentration rules have the property that concentration goes down under aggregation. The corresponding property of the sensitivity function is that it is subadditive. Part of the modern analysis of the cell suppression problem showed that the improvement limitation rules are also subadditive. They have the added advantage that they are motivated by the estimates that can be formed about the enterprises by various observers, including other enterprises. The same analysis of the concentration rules shows that they provide a variable amount of protection for the enterprises. The newer form can meet both objectives, while the older form can only meet the objective of subadditivity. One would expect that use of the older form would be very hard to justify, given both the analysis and the experience with the newer form. The forms displayed have restricted their last coefficient to be minus one. This allows both upper and lower bounds to be placed on the sensitivity of aggregations. We have

$$S(x+y) \leq S(x) + S(y)$$

$$S(x+y) \geq S(x) - T(y).$$

■ Aggregations

In the documentation of the USBC System, we learn that some cell aggregation is done after the main tabulation phase. These cell aggregations are based on the largest two contributors to the cells. We would expect problems to be present if the identity of the third and higher contributors is lost. For example,

100	100
A	Total

+

20	40	40	100
Etc(A)	B	C	Total

=

60	40	100	200
Etc(A,B)	C	A	Total

does not have the two largest contributors more than 75 percent, as we have lost track of the pieces of A. If we keep all the identities, we have

100	100
A	Total

+

20	40	40	100
A	B	C	Total

=

40	40	120	200
B	C	A	Total

and the two largest contributors are more than 75 percent. This would be an error for the 2-75%-concentration rule given above. Other values would be needed to illustrate the error for other parameter values. USBC does not publicly specify the parameters of its rule. If these two cells were in a row of a table, we would like to know whether the aggregation is sensitive or not. If the aggregation is sensitive, then we would require additional suppression within the row to protect the respon-

dents, as the cells cannot act as complements for each other. This is an example of the influence of the common ownership.

R1	X	X	-	-

If this were an isolated row, we could just collapse the classifications. This is commonly called category rollup. When this row is part of a larger table, we have the two problems that we cannot collapse the classifications in the other rows, and we have allowed the values to be recovered by subtraction.

T	C1	C2	C3	C4
R1	X	X	-	-
R2	-	-	-	-
R3	-	-	-	-
R4	-	-	-	-

We avoid the problems by having additional suppressed cells, so that we have

T	C1	C2	C3	C4
R1	X	X	-	-
R2	X	X	-	-
R3	-	-	-	-
R4	-	-	-	-

We now have collapsed classifications in both some of the rows and some of the columns. A full description of the collapsed classifications is not usually provided. The user is left to notice that the total of rows R1 and R2 in column C1 is determined exactly. The incorrect rule that two suppressions in every row and column will do the job is sometimes suggested. The rule must be satisfied, but it can lead to errors. The entry in row R2 and column C3 below is available with only a small amount of arithmetic. If we had documented the various collapsed classifications, we would avoid this mistake. We would have been led to notice that the total of the cells in columns C1 and C2 of row R2 is known (similarly, for rows R3 and R4 of column C3).

T	C1	C2	C3	C4
R1	X	X	-	-
R2	X	X	X	-
R3	-	-	X	X
R4	-	-	X	X

The example has been displayed so often that one could gain the impression that it is the only known problem. It is so trivial that it has many easy explanations. It should be called a silly mistake that should never be seen in practice. It can hide easily, as the following permuted example shows.

T	C1	C3	C2	C4
R1	X	-	X	-
R3	-	X	-	X
R2	X	X	X	-
R4	-	X	-	X

■ Metropolitan Areas

Metropolitan areas are the major example of nonhierarchical classifications. Geographical classifications are usually organized with a progression of refinements of larger areas into smaller areas. We will call these region, county, and place. (Local legal history often assigns different terminology, so we must treat these as definitions of notions.) The regions are mutually exclusive and exhaust some larger geographical area. The counties are mutually exclusive and exhaust the regions, and the places are mutually exclusive and exhaust the counties. For classification purposes, all the areas are assigned codes. The result would be a code with an "rrccpp" structure. This simple structure does not deal with the practical reality of metropolitan areas. A metropolitan area is a grouping of adjacent urban places, which overlaps a collection of counties with urban and rural places. There is also the implicit grouping of rural places defined by exclusion from the urban metropolitan area. The urban and rural metropolitan areas are mutually exclusive and exhaust their regions, just like the counties. The places could be assigned a locality

code within their metropolitan areas to construct an "rrmmll" structure. As an example, we have county A with places Ax, Ay, and Az; B with Bx, By, and Bz; and C with Cx, Cy, and Cz. Depending upon which coding structure we use, we could construct a table of county and place by metropolitan area

	Total	Urban	Rural
R	#r	#u	#{r-u}
A	#a	#ax	#{ay+az}
Ax	#ax	#ax	0
Ay	#ay	0	#ay
Az	#az	0	#az
B	#b	#{bx+by}	#bz
Bx	#bx	#bx	0
By	#by	#by	0
Bz	#bz	0	#bz
C	#c	#cx	#{cy+cz}
Cx	#cx	#cx	0
Cy	#cy	0	#cy
Cz	#cz	0	#cz

or a table of metropolitan area and locality by county

	Total	A	B	C
R	#r	#a	#b	#c
Urban	#u	#ax	#{bx+by}	#cx
Ax	#ax	#ax	0	0
Bx	#bx	0	#bx	0
By	#by	0	#by	0
Cx	#cx	0	0	#cx
Rural	#{r-u}	#{ay+az}	#bz	#{cy+cz}
Ay	#ay	#ay	0	0
Az	#az	#az	0	0
Bz	#bz	0	#bz	0
Cy	#cy	0	0	#cy
Cz	#cz	0	0	#cz

In the practical world, we often see a *unified* code of "rrccmmpp," which is a single code but represents two dimensions. We never see the above displays with their many zeros and rarely see the urban and rural subtotals for counties or the rural subtotal for the region. The unified code addresses the important practical geographical issue that the urban or rural status is not well represented by counties. The result is a nonhierarchical classification structure. Other examples arise when the historical continuity of *ad hoc* code groupings is maintained across a classification revision, as has happened with the Standard Industrial Classification revisions. Nonhierarchical classifications do not form networks as used with some cell suppression programs.

■ Three-Way Example

Live data from the Manufacturing Energy Consumption Survey (MECS) for 1991 was used to do a comparison of automated cell suppression systems. MECS is sponsored by the Energy Information Administration (EIA), with the survey operation carried out under contract by the USBC. All access to confidential microdata was on USBC premises, and all working tables were only available there before they were destroyed. The microdata were:

SIC	Geo	FuTy	CoId	Data

SIC--Standard Industrial Classification
- type of business

Geo--Geography
- location of business

FuTy--Fuel Type
- fuel type used by business

CoId--Company Identification
- ownership of business activity

Data--Energy Consumption
- amount of fuel used
- confidential

Three systems were used in this comparison. The first was the USBC System developed by USBC. This system is intended to be operated by its developers and was for this study. The second was CONFID from Statistics Canada. It was made available to some U.S. agencies under restricted conditions with no support. It is

intended to be operated by its end users. The third was Automated Cell Suppression (ACS), which is a newly developed follow-on to CONFID, independent of Statistics Canada. It is intended to be operated by its end users. The USBC System is based on network theoretic notions that apply to two dimensions with some restrictions. A third dimension is treated as many independent one-way problems and may not be numerically secure. CONFID and ACS use general simplex notions that apply to any number of dimensions. CONFID is restricted to either two or three dimensions by its implementation. The initial testing and comparisons were done with a single three-way table from the MECS publication. This allowed some initial specification errors to be detected and fixed. Cross auditing of the systems was possible with this problem. The USBC System has a postprocessing stage to detect a restricted class of residual disclosures. Its audit capability is rarely used, as it is too slow to operate. CONFID and ACS provide full audit capabilities.

USBC
- 1 inadvertent approximate residual disclosure.
- oversuppression as 76 of 876 suppressions released.

CONFID
- a test file produced a spurious failure.

ACS
- default testing of miscellaneous aggregation inadequate.

The oversuppression was detected by using the USBC System suppression pattern as a starting point for ACS, which then reported unused presupplied suppressions. The failure of CONFID was to report, by its audit component, as nonadding the output of its suppression component. This should not happen, as it was, in turn, the output of the tabulation component. The adjustment component reported the table as exactly additive. An audit by ACS agreed and further showed that the suppression component had introduced spurious isolated suppressions. This appears to be related to defective maintenance of CONFID at Statistics Canada, as the two suspect components had received attention for a differing issue. The ACS aggregation-testing default was inadequate for the high incidence of common ownership across the fuel type classification. Setting the testing parameter to other than the default value resolved the issue.

■ Five-Way Example

The MECS data have two additional classification variables.

SIC	Geo	FuTy	PrTy	CoTy	CoId	Data

SIC--Standard Industrial Classification
- type of business

Geo--Geography
- location of business

FuTy--Fuel Type
- fuel type used by business

PrTy--Production Type
- energy production type of business

CoTy--Consumption Type
- energy consumption type of business

CoId--Company Identification
- ownership of business activity

Data--Energy Consumption
- amount of fuel used
- confidential

There are three Production Types and two Consumption Types for six components of energy use. One of the components is out of scope for MECS and defined to be zero. The publication is complicated, as only four selected three-way subtotals are published for these data. The subtotals are standard groupings of components that are used in energy analysis. In the publication, the four subtotals are Tables A1, A3, A4, and A5. Table A5 is a subset of Tables A1 and A4, and Table A3 is a subset of Table A1.

USBC
- 75 inadvertent exact residual disclosures
- structure represented by inequalities

ACS
- five-way reduced to four-way nonhier-

archical to lower redundancy

CONFID could not process the five-way structure. In earlier runs, the USBC System had more suppressions but fewer residual disclosures. The system had run as intended and had no disclosures on the separate three-way problems. The conclusion was that the disclosures were the result of not representing the five-way structure. The protection was the result of alignment of suppressions within the larger system and was lower when there were fewer suppressions.

Both CONFID and ACS offer multiple objective functions. The "Constant" objective function attempts to minimize the count of suppressed cells and is named after the constant weights applied to the cells. The "Size" objective function attempts to minimize the value of suppressed cells and is named after the objective function coefficients, which are the cell sizes. Minimizing the count of cells often suppresses large cells. Minimizing the value of suppressed cells often suppresses many small cells. The "Digits" objective function attempts to balance between these two by minimizing an entropy suggested by information theory and is named after objective function coefficients, which are a modified logarithm of the cell values. Cells were classified as marginal or internal cells to help with a multiple counting problem. Whenever a marginal cell is suppressed, some of its internal cells must also be suppressed. The total value for internal cells can be compared to the grand totals given below. The values will not match the publication, as the processing was done on absolute values,

"Constant" Objective Function for ACS

		All Cells	Margin Cells	Internal Cells
Jointly	Count	1317	1190	127
	Value	226274	221960	4313
Table A1	Count	497	364	133
	Value	103852	96794	7057
Table A3	Count	102	90	12
	Value	20958	19845	1113
Table A4	Count	355	263	92
	Value	67711	63498	4213
Table A5	Count	363	248	115
	Value	33750	30551	3199

"Digits" Objective Function for ACS

		All Cells	Margin Cells	Internal Cells
Jointly	Count	1373	1198	175
	Value	171449	165435	6014
Table A1	Count	479	335	144
	Value	71494	66718	4776
Table A3	Count	121	100	21
	Value	23918	21686	2231
Table A4	Count	319	210	109
	Value	39002	35750	3252
Table A5	Count	454	300	154
	Value	37034	33251	3782

"Size" Objective Function for ACS

		All Cells	Margin Cells	Internal Cells
Jointly	Count	1812	1590	222
	Value	105016	99134	5882
Table A1	Count	609	385	224
	Value	30821	24893	5927
Table A3	Count	159	138	21
	Value	19046	16528	2517
Table A4	Count	464	292	172
	Value	24214	20363	3851
Table A5	Count	580	379	201
	Value	30934	27570	3364

USBC System

		All Cells	Margin Cells	Internal Cells
Jointly	Count	2664	2267	397
	Value	90073	82534	7538
Table A1	Count	759	382	377
	Value	16739	10839	5900
Table A3	Count	163	132	31
	Value	13008	10370	2638
Table A4	Count	913	539	374
	Value	40561	33974	6586
Table A5	Count	829	463	366
	Value	19763	14863	4900

Grand Total	
Jointly	18804
Table A1	17662
Table A3	3503
Table A4	15301
Table A5	10765

as some net values of electricity or steam consumption can be negative. The publication also has some values from other EIA, but non-MECS, sources. The results for the USBC System are similar to the results for ACS with the "Size" objective function, as they use the same coefficients. It is unknown how many additional cells, or their values, would have to be suppressed to correct the identified problems.

■ Conclusions

MECS is a complex problem of moderate size. The five-way structure is more complex than is typical of economic surveys.

All the systems operate more quickly than analysts can review their output, so the variations in execution cost are not a qualitative comparison attribute. The USBC system is not numerically secure in three dimensions and produced exact residual disclosures for the five-dimensional MECS tables. It is prone to over suppression as demonstrated by example in this study. There is no usable audit facility. CONFID suffered from spurious failures and is limited to two or three dimensions. ACS completed these tests and demonstrated a full audit capability.

The value of the ability to independently verify the successful operation of a suppression program was strongly illustrated by the problem encountered with CONFID. This lesson was reinforced by the need to verify that the USBC system had operated as intended for the three-way tables when it produced inadvertent residual disclosures in the five-way table.

See "Report on Statistical Disclosure Limitation Methodology" for an overview of the topic and an extensive bibliography. See "Report on EIA-Census Evaluation of Disclosure Limitation Methods" for a full report of this study.

■ References

OMB (1994), "Report on Statistical Disclosure Limitation Methodology," *Statistical Policy Working Paper 22*, Office of Management and Budget, Washington, D.C.

OSS (1996), "Report on EIA-Census Evaluation of Disclosure Limitation Methods," Office of Statistical Standards, Energy Information Administration, Washington, D.C.

* Gordon Sande
Sande and Associates, Inc.
600 Sanderling Court
Secaucus, NJ 07094