BUREAU OF THE CENSUS
STATISTICAL RESEARCH DIVISION
RESEARCH REPORT SERIES
No. RR-92/09

AN OVERVIEW OF DISCLOSURE PRINCIPLES

by

Colleen M. Sullivan
U.S. Bureau of the Census
Statistical Research Division
Washington, D.C. 20233

Report issued: September 22, 1992

An Overview of Disclosure Principles

Colleen M. Sullivan

1. INTRODUCTION

The Bureau of the Census operates under Title 13 of the U.S. Code, which prohibits the Bureau from making "any publication whereby the data furnished by any particular establishment or individual under this title can be identified." This rule prohibits the Bureau from publishing a summary table that enables a data user to derive detailed information about an individual respondent. To ensure our tables do not violate disclosure rules implied by Title 13, they must first be subjected to an analytical procedure referred to as disclosure analysis. Disclosure analysis begins with the simple principle that we must not directly publish data received from individuals who respond to our Economic surveys and censuses.

This paper is organized as follows: A description of sensitive data and of the cell suppression method that is used to protect the sensitive data in publications appears in Section 2. Section 3 presents a description and discussion of the use of complementary suppressions. Section 4 explains how to estimate a range of feasible values for all suppressed cells. The two types of primary suppression rules used at the Census Bureau are examined in Section 5. Section 6 addresses the cost of suppressions schemes and a summary appears in Section 7.

2. SENSITIVE DATA

The Economic Divisions have the responsibility to collect a wide range of data and to publish these data without violating confidentiality laws. Normally, economic data is published by geography and standard industrial classification (SIC) codes. For example, Table 1 shows state level data for various types of food stores.

SIC	Number of Establishments	Value of Sales
54 All Food Stores . . .	347	$200 900
541 Grocery	333	196 000
542 Meat and Fish .	11	1 500
543 Fruit Stores . . .	2	2 400
544 Candy	1	1 000

Table 1. Typical Data Table

This table shows that only one establishment reported candy store sales for this state. If this table were published, any data user would know the establishment's precise sales value. Also, this table shows only two establishments reporting fruit store sales. Either of these two establishments, knowing their own sales figure, would be able to calculate the other establishment's precise sales figure. Thus, publishing this table would result in a disclosure, violating Title 13. Values such as these are considered sensitive, and must not be published (i.e, disclosed). Values which would disclose an individual's or establishment's data are termed sensitive.

Sensitive Data Values:

A data item which a data user could utilize to calculate another individual's or establishment's data. Sensitive data values must not be disclosed.

One way to prevent the identification of sensitive values is to simply not publish the values. When we publish this table, we would replace the sensitive data values with a "(D)". Table 2 shows a publishable table where the sensitive data values have been suppressed.

SIC	Number of Establishments	Value of Sales
54 All Food Stores	347	$200 900
541 Grocery	333	196 000
542 Meat and Fish .	11	1 500
543 Fruit Stores . . .	2	(D)
544 Candy	1	(D)

Table 2. Protected Respondent Data

This disclosure avoidance technique is referred to as cell suppression. (Note that although a data value may be sensitive, the corresponding number of establishments is not, and therefore is never suppressed.)

Cell Suppression:

A disclosure avoidance technique in which the sensitive data in the publication is replaced with a "(D)".

3. COMPLEMENTARY SUPPRESSIONS

If we only suppress sensitive data, users could frequently derive the values from non-sensitive data because most data items are published in additive tables. Notice that the suppressed value in Table 3 can be derived by subtracting the non-suppressed interior cell values (5,413 and 61,252) from the row total (84,842). By performing this calculation, we determine that the suppressed data value must be 18,177.

	State	MSA 1	MSA 2	NON-MSA
SIC Total	173 536	14 566	45 105	113 865
SIC 1	84 842	5 413	(D)	61 252
SIC 2	43 588	1 377	20 146	22 065
SIC 3	45 106	7 776	6 782	30 548

Table 3. Additive Table

Therefore, to fully protect the suppressed sensitive data value, additional data values must be suppressed. These new suppressed cells are referred to as complementary suppressions.

Complementary Suppressions:

Suppressions which prevent a user from deriving a sensitive data value from additive table relationships.

Table 4 presents a set of complementary suppressions that protects the sensitive data value. Note the "(C)" notation is used only in this documentation; a "(D)" would appear in the actual publication.

	State	MSA 1	MSA 2	NON-MSA
SIC Total	173 536	14 566	45 105	113 865
SIC 1	84 842	5 413	(D)	(C)
SIC 2	43 588	1 377	(C)	(C)
SIC 3	45 106	7 776	6 782	30 548

Table 4. A Suppression Scheme

We must be certain that no suppressed values can be derived exactly. It is rarely sufficient to merely look at a table and determine that the complementary suppression scheme fully protects all suppressed values. Often a two dimensional table seems to have an adequate number of complementary suppressions, but mathematical manipulations reveal a suppressed data value.

Consider the following table where each cell with a letter is being suppressed. We ask: Can we determine the value in row 3, column 3 (cell k)?

	Total	Column 1	Column 2	Column 3	Column 4
Total	510	100	100	160	150
Row 1	155	25	**a**	40	**b**
Row 2	125	**e**	20	**f**	30
Row 3	150	30	**c**	**k**	**d**
Row 4	80	**g**	10	**h**	20

At first, it certainly seems that there is a sufficient number of suppressions to protect the value of k. However, we can determine the value of k by using some basic algebraic techniques.

Observe the following:

Column 2 =>	100 = a + 20 + c + 10	=> a + c = 70	(1)
Column 4 =>	150 = b + 30 + d + 20	=> b + d = 100	(2)
Row 1 =>	155 = 25 + a + 40 + b	=> a + b = 90	(3)

Adding (1) and (2) yields a + b + c + d = 170
and subtracting (3) -(a + b = 90)
---
yields c + d = 80

Now observe :

Row 3 => 150 = 30 + c + k + d => k = 120 - (c + d).

Substituting in c+d=80 from the above calculation yields k = 120 - 80 = 40. Thus, we have determined that k = 40. To fully protect all suppressed values, more values must be suppressed. We must then recheck the table to ensure no values can be derived through algebraic techniques.

4. FEASIBLE RANGES

Although we ensure that data users cannot estimate a suppressed data value exactly, a range of feasible values for any suppressed cell can be estimated. For a simple example, consider Table 5 where all the values have been suppressed.

18	10	8
7	D1	D2
11	D3	D4

Table 5.

Knowing that the table is additive and that all values are non-negative, we ask: "What is the smallest value we can assign to D1 and still have the table be additive?"

> If we let D1=0, then D2 must be seven since D1+D2 = 7.
> Then D3 must be ten since D1+D3=10.
> Therefore, D4 must be one.
> Thus we can say a lower bound for D1 is zero.

Now we ask: "What is the largest value we can assign to D1 and still have an additive table?"

> If we let D2 =0, then D1 must be seven since D1+D2=7.
> Examine the other equation with D1: D1 + D3 = 10.
> In this equation D1 cannot equal 10 because D1+D2=7 tells us the most D1 can be is seven.
> Thus, we can say that an upper bound for D1 is seven.

The feasible values for D1, in this example, fall in the range $0 \leq D1 \leq 7$. We could also calculate ranges in this manner for all other suppressions in this table. However, not all suppressed tables are as simple as presented here. Therefore, data users rely on linear programming techniques to determine the feasible ranges for suppressed cells.

5. PRIMARY SUPPRESSION RULES

Table 1 showed two obvious disclosures -- only one or two firms contributed to a cell. A not so obvious disclosure occurs when more than two firms contribute to a data cell, but one firm is able to estimate the data for another firm very closely. This type of disclosure is detected through application of a primary suppression rule. A cell that cannot be published because it fails the primary suppression rule is called a primary suppression. There are two types of primary suppression rules used at the Census Bureau, the n-k rule and the p% rule. The n-k rule is aimed at protecting the value of each company from a coalition of (n-1) other companies in the cell. This rule states that a cell must be suppressed if the largest n respondents in the cell make up at least k% of the total cell value. The p% rule is aimed at protecting the largest, and therefore all, company values in a given cell from upper estimation to within p%. In the following discussion, the p% rule will be used.

Primary Suppression:

A cell that cannot be published because it fails either the n-k rule or the p% rule.

To illustrate the p% primary suppression rule,

Let T = the total value of a given cell,
L = the value of the largest contributor to the cell,
S = the value of the second largest contributor to the cell, and
p = the percentage of protection required.

Then $R = T - L - S$ is the total value of the remaining contributors to the cell.

The p% rule states that a cell must be suppressed if $R < (p/100)*L$. The value of p, itself, is considered sensitive and is not revealed to anyone outside the Census Bureau. For example, consider the cell (18,177) in Table 6. Suppose it is composed of L=$17000, S=$1000, and R=$177. Also suppose the value of p is 15.

	State	MSA 1	MSA 2	NON-MSA
SIC Total	173 536	14 566	45 105	113 865
SIC 1	84 842	5 413	D (18 177)	61 252
SIC 2	43 588	1 377	20 146	22 065
SIC 3	45 106	7 776	6 782	30 548

Table 6. Additive Table

The p% rule indicates this cell is a primary disclosure since 177<(15/100)*17000=2550. If we were to publish this cell, most people could not determine much about the data for the largest contributor. However, the owner of the second largest contributor knows his

sales are $1000, and he could subtract that number from the published total to derive that the sales for the largest contributor were less than $17,177, which is within 15% (actually within 2%) of the true value. Under the p% rule with p=15, this would be disclosing too much information about the largest contributor, and we would suppress this cell. Therefore, a "(D)" would appear in the published table instead of the value 18,177.

Recall, from Section 4, data users are able to calculate a range of feasible values for any suppressed cell. However, when choosing complementary suppressions for some primary suppression with true value X, we ensure that it cannot be estimated within a smaller interval than $X \pm B$ where B is the amount of lower and upper protection required by X. The p% suppression rule implies that $B=(p/100)L-R$. That is, we need to choose complementary suppressions having a minimum value of $(p/100)L-R$. This is the minimum value needed to protect the sensitive data value by p%. (Note the n-k rule implies a different value for B.)

Using the previous example in this section, the sensitive data value (18,177) must be protected by a value of at least $(15/100)*17000-177=2373$. In other words, the data values chosen to be in the suppression scheme must be at least 2373 in value. If it is not possible to accomplish the protection by selecting only one cell in a row or column, then a set of cells totalling 2373 must be chosen in the row or column to serve as complementary suppressions.

6. SUPPRESSION SCHEME COST

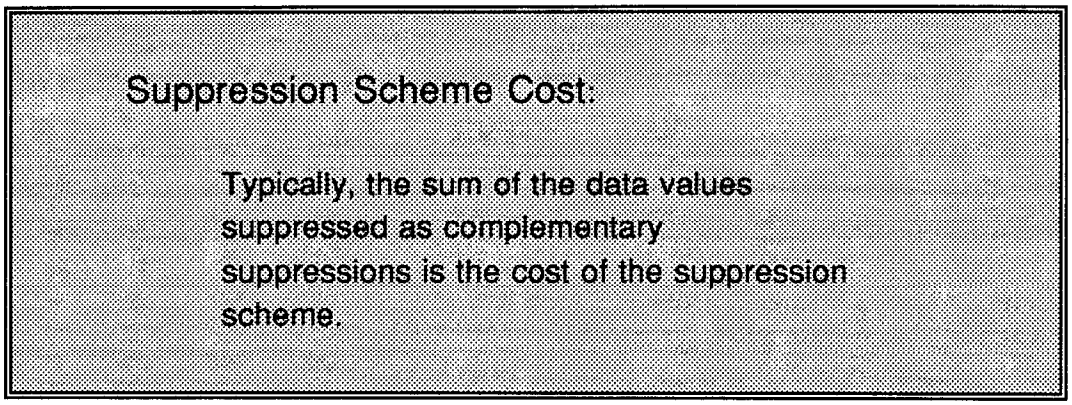
Table 4 in Section 3 showed one complementary suppression scheme that protected the sensitive data value. However, this is not the only scheme that would have protected the sensitive data value. We could have chosen to suppress the values shown with a "(C)" in Table 7.

	State	MSA 1	MSA 2	NON-MSA
SIC Total	173 536	14 566	45 105	113 865
SIC 1	84 842	(C)	(D)	61 252
SIC 2	43 588	1 377	20 146	22 065
SIC 3	45 106	(C)	(C)	30 548

Table 7. An Alternative Suppression Scheme

The sum of the complementary suppressions in Table 4 is 103,463, while the sum of the complementary suppressions in Table 5 is 19,971. (Both suppression schemes ensure that the sensitive data value is protected by the required 2373 units as mentioned in Section 5.) Less total data value is suppressed by the complementary suppression scheme of Table 5, and thus it is the preferred scheme.

The objective in applying complementary suppressions is to ensure the protection of the sensitive data value at minimum cost. Note that this requires assigning a cost of suppression to each data cell. Usually, the original data value that would have appeared in the publication is assigned as the cost. By minimizing the cost incurred through complementary suppressions, the greatest amount of usable data is provided.



7. SUMMARY

We have seen that disclosure analysis begins with the simple principle that we must not directly publish data received from individuals who respond to our Economic surveys and censuses. The simplest and most obvious of all sensitive data is that in which only one or two firms contribute to a particular cell. Obviously, these values must be suppressed in any publication. Next, we saw that through application of the primary suppression rule, whether it be the n-k rule or the p% rule, other sensitive cells may exist and must also be suppressed. Because most data appear in additive tables, relations exist which require the use of complementary suppressions to protect the already suppressed sensitive data. Still, we do not want any respondent's value estimated exactly or "too closely." Therefore, we must ensure that these complementary suppressions provide the required amount of protection for the sensitive cells. Finally, there is the matter of using

mathematical manipulation on tables to derive suppressed values. If a value is derived, whether it was a primary or complementary suppression, we have violated our confidentiality law. Thus, we must ensure that no suppressed values are derivable. This paper has merely reviewed the disclosure principles that must be enforced when publishing our tables.

FURTHER READINGS

There are a number of methods available which prevent compromising primary suppressions. These disclosure avoidance techniques include rounding, perturbation, and cell suppression, and are discussed in the following papers:

Cox, L.H. (1975), "Disclosure Analysis and Cell Suppression ," *Proceedings of the American Statistical Association, Social Statistics Section.*

Cox, L.H., and Ernst, L.R. (1982), "Controlled Rounding," *INFOR*, 20, 4, 423-432.

Cox, L.H., Fagan, J.T., Greenberg, B.V., and Hemmig, R.J. (1986), "Research at the Census Bureau into Disclosure Avoidance Techniques for Tabular Data," *Proceedings of the American Statistical Association, Survey Research Methods Section.*

Cox, L.H., and George, J.A. (1989), "Controlled Rounding for Tables with Subtotals," *Annals of Operations Research*, 20, 141-157.

Cox, L.H., McDonald, S., and Nelson, D. (1986), "Confidentiality Issues at the United States Bureau of the Census," *Journal of Official Statistics*, 2,135-160.

The Bureau currently utilizes network flow methodology as a means of choosing complementary suppressions for economic surveys and censuses. This methodology is discussed in the following papers:

Cox, L.H. (1980), "Suppression Methodology and Statistical Disclosure Control," *Journal of the American Statistical Association*, 75, 377-385.

Cox, L.H., Fagan, J.T., Greenberg, B.V., and Hemmig, R.J. (1986), "Research at the Census Bureau into Disclosure Avoidance Techniques for Tabular Data," *Proceedings of the American Statistical Association, Survey Research Methods Section.*

Gusfield, D. (1984), "A Graph Theoretic Approach to Statistical Data Security," Department of Computer Science, Yale University, New Haven.

Kelly, J.P., Golden, B.L., and Assad, A.A. (1992), "Cell Suppression: Disclosure Protection for Sensitive Tabular Data," *Networks*, 22, 397-417.

Sullivan, C.M. and Rowe, E.G. (1992), "A Data Structure and Integer Programming Technique to Facilitate Cell Suppression Strategies," *American Statistical Association, 1992 Proceedings of the Section on Survey Research Methods*, to appear.

Sullivan, C.M. and Zayatz, L. (1991), "A Network Flow Disclosure Avoidance System Applied to the Census of Agriculture," *American Statistical Association, 1991 Proceedings of the Section on Survey Research Methods.*

Sullivan, C.M., and Zayatz, L. (1992), "A Disclosure Avoidance System Using Network Methodology for the Census of Agriculture," SRD Census Confidential Research Report Series, No. CCRR-92/02, Bureau of the Census, Statistical Research Division, Washington, D.C. 20233.

ACKNOWLEDGEMENTS

I gratefully acknowledge the help of those who reviewed and provided motivation for this paper, especially Alan Saalfeld, Robert Jewett and Laura Zayatz. Special thanks are also due to Dennis Shoemaker, Bill Wester and Peggy Allen for comments on an earlier version.