RESEARCH REPORT SERIES
(Disclosure Avoidance #2016-05)

Predicting Complementary Cell Suppressions

Given Primary Cell Suppression Conditions

Phyllis Singer and Nelson Chung

Center for Disclosure Avoidance Research
U.S. Census Bureau
Washington DC 20233

Report Issued: September 30, 2016

Abstract

The U. S. Census Bureau uses cell suppression methodology as the primary disclosure avoidance methodology for various economic surveys and the Economic Census. A research group has been involved in updating cell suppression processing for the 2017 Economic Census. Previous results show that ordering primary cell suppression affects the output of the cell suppression program. This paper describes two of the many research projects for improving cell suppression at the Census Bureau. One project was to use modelling to predict what happens for a particular primary cell in the cell suppression routine. The other project was to estimate empirically the long-term average number and sum of secondary suppressions for randomly ordered primary suppressions.

Key Words: Disclosure Avoidance, Cell Suppression

Table of Contents

1 Introduction

The United States Census Bureau pledges to respondents that the data it collects will be used for statistical purposes only. Further, the Census Bureau pledges not to release data that will identify respondents or their attributes. For economic surveys, respondents may be firms or establishments.

The Census Bureau publishes a great many tables about the economy. Economic respondents' privacy currently is safeguarded through two main methods: suppressing sensitive table cells and noise infusion. The Census Bureau currently uses cell suppression for the following economic programs.

- Annual Capital Expenditures Survey
- Annual Survey of Manufactures
- Business R&D and Innovation Survey
- Economic Census of the United States
- Manufacturing Energy Consumption Survey

A table cell is considered sensitive if releasing that data would allow estimating a single contributor's value too closely. This situation occurs when there are very few contributors, or when one or two large contributors dominate the aggregate statistic. Such cells are called primary suppressions. Secondary suppressions are additional cells that must be suppressed so that the primary suppressions cannot be estimated from the other cells in the table together with the table margins. Current practice (Federal Committee on Statistical Methodology 2005, 61) is to detect sensitivity by using the p-percent rule.

This paper concentrates on the problem of secondary suppressions. One problem is developing models that predict, for each primary suppression, the number and sum of protections for secondary suppressions (called, for brevity, the sum of secondary suppressions) from characteristics of the underlying data. Another problem is estimating the overall average number and sum of secondary suppressions from repeated runs of the research data suppression program. The repeated runs had, as one of their inputs, different permutations of the primary suppressions. The modeling was designed to produce local estimates for each primary suppression, and the empirical study was designed to supply global measures.

2 Literature Review

Cox (1975, 380) introduced cell suppression as a disclosure avoidance method suited for demographic and economic censuses in which (1) there are "many levels of aggregation" and (2) data are "inhomogeneously distributed." He made three observations about suppression. First, a table with the same number of rows and columns and with one complementary cell suppression requires at most three additional suppressions (380). Second, if the number of rows and columns are not equal, then the number of complementary suppressions equals the number of rows or the number columns, whichever is greater (380). Finally, the number of distinct suppression schemes equals the factorial of the number of complementary suppressions (380).

Cox, Fagan, Greenberg, and Hemmig (1986, 388) considered cell suppression, rounding, and perturbation as three techniques with a unifying mathematical structure. They found exact estimation of primary suppression cell values can be avoided by placing it in a circuit of suppressed cells (391), and that an agency can calculate the range that contains the cell value derived from solving a system of linear equations by employing a concept of “capacitated network flow” (391). While perturbation requires complementary perturbations as well, complementary suppressions hold an advantage over complementary perturbations because one complementary suppression can protect multiple primary suppressions (393).

Massell (2002) and Wang (2013) identified advantages of linear programming (LP) techniques over network flow models. Massell (2002) compared different cell suppression methods. There is less oversuppression when using LP techniques than when using network flow models. Integer programming (IP) techniques offer the optimal solution to the cell suppression problem but may be too slow to implement for large tables. Wang (2013) demonstrated how LP overcomes problems that network flow models have in handling complex tables. Steel et al. (2013) elaborated further on the processing, describing the setup of the LP as involving:

(1) searching for a pattern to protect the cells according to the p-percent rule

(2) creating “super cells” to ensure that unions of suppressed cells do not violate the p-percent rule

(3) modifying the model to accommodate the supercells and obtaining a solution that minimizes the cost in suppression

Wang (2016) demonstrated that in most cases, ordering the primary suppressions based on associations and protection levels resulted in somewhat fewer complementary cells suppressed in comparison to random orders.

3 Methodology and Results

Below we provide the methodology and results for the two parts of our research. The first section discusses the modeling study, while the second discusses the empirical study of suppressions for a subsector of the 2012 Economic Census.

3.1 Model Study

This section is split into two parts. The first discusses the model used for predicting the number of complementary suppressions for a given primary suppression. The second part discusses the model for predicting the total value contained in complementary suppressions for a given primary suppression.

3.1.1 Model for Number of Complementary Suppressions for a Primary Suppression - Methodology

It would be useful to know in advance what we can expect to happen for a particular primary (P) in cell suppression. A strong positive correlation is thought to exist between the value in the cell and the protection requirement, because the requirement is based on the p% rule. However, this is not always the case, because the p% rule has a remainder term. Therefore, a precise prediction can drive better processing order. The prediction can also be used in review, which is important because the Census Bureau has long pushed the idea of focusing the data review on cases with the greatest impact on the final product. The disclosure avoidance process has been excluded for lack of a way to generate review cases.

Several major obstacles to prediction exist. The cell suppression process only directly addresses a small percentage of the actual primaries, while the rest are determined to be protected by processing already executed. The raw input data may contain duplicates. If P1 and P2 are duplicates, then the pattern generated for whichever P is encountered first protects its duplicate. This is because two duplicates essentially name the same cell. Order matters, because cells processed towards the end can be protected by cells already marked as a complementary suppression with no extra cost. Once a cell is marked as a complementary cell it can be used to protect primary cells that occur later. Company reports force one to classify vertical complexes together. A vertical complex is a company that spans different NAICS[1] or geographical levels. We can gain efficiency by selecting Ps that generate Cs that protect the entire complex.

We created a special training data set from a hierarchical 2012 Economic Census table for Sector 72, Accommodation and Food Services, where the update process was disabled. This means that Cs were not recorded for subsequent problems. We obtained a list of complementary cell suppressions for each P as if it were the first P processed.

We analyzed the Ps that accounted for their context in the tables being processed. We asked the following questions:

- How many tables was the P in?
- Was it a margin in those tables?
- Were neighboring cells already suppressed?
- How far was it from the grand total?
- What was its protection requirement?

We propose the following model to predict the suppression pattern for a P:

$$\textbf{(1)} \quad \boldsymbol{ln(NUMC) = \beta_o + \beta_1 PROT + \beta_2 PROT^2 + \beta_3 NUMREL + \beta_4 CAPSIB_{AVG} + \beta_5 DEPTH + \beta_6 NUMCOMP1 + \beta_7 VAL + \beta_8 NUMTOT + \varepsilon}$$

Where:

NUMC – Number of Complementary Suppressions. Each primary suppression (P) requires a number of complementary suppressions to protect it. This is the number of complementary suppressions (C) required, if that P is done first.

[1] By NAICS we refer to the code assigned to an industry in the North American Industry Classification System.

PROT– Protection Requirement. The amount by which the cell fails the sensitivity rule. It represents how far from the true value the bounds on the cell need to be. Any estimate of the cell by a competitor can fail by as much as p%.

NUMREL – Number of relations. The number of tables the cell is in, which lines up one-to-one with the instances of relations).

CAPSIB_AVG – Average Capacity of Siblings. For each relation we have the sum of suppressed value (excluding the reference P). Siblings is a misnomer; it could also have parent or child. Capsib_avg is the average over all the relations the cell is in.

DEPTH – Depth. The number of steps away from the grand total.

NUMCOMP1 – Number of times the company one report appears in other cells (id1 and value). Could be as company one or this could be expanded to where the report appears as company 2. This is to account for vertical structure.

VAL – Value. The value of the protected cell.

NUMTOT – Number of times cell appears as a total.

We use Poisson regression modelling to predict the number of complementary suppressions each primary suppression induces. Since the number of complementary suppressions is non-negative, discrete count data, we expect the data to exhibit a Poisson distribution.

3.1.2 Model for Number of Complementary Suppressions for a Primary Suppression - Evaluation and Results

We ran the regression using the SAS® GENMOD procedure.

The scale parameter, which relates the mean to the variance, equals one. Poisson modelling assumes the first two moments equal. To test this assumption, we used the following diagnostic in the SAS® GENMOD procedure:

Table 3.1.1 Deviance and Pearson Residuals for Number of Complementary Suppressions for a Primary Suppressions, $\sigma^2 = \mu$.

Criterion	DF	Value	Value/DF
Deviance	5463	11984.1585	2.1937
Scaled Deviance	5463	11984.1585	2.1937
Pearson Chi-Square	5463	15066.5373	2.7579
Scaled Pearson X^2	5463	15066.5373	2.7579

The Deviance is computed using the following formula:

$$(2) \qquad D\left(\hat{\mu}_s, \hat{\mu}, \hat{\phi}\right) = \sum_i^n 2\{l_i(\hat{\mu}_{si}, \hat{\phi}) - l_i(\hat{\mu}_i, \hat{\phi})\}\ .$$

That is, the Deviance equals the sum of two times the difference between the log likelihood of the saturated model and the fitted model. The ϕ represents a scale parameter that relates the variance to the mean.

The Deviance statistic is the sum of the squared deviance residuals. Both the Pearson Chi-square and the Deviance are significantly greater than unity, meaning the variance is greater the mean, i.e., data is *overdispersed.* The model requires rescaling based on either indicator. We choose the Deviance statistic because of the availability of an R^2 computation for the model in the literature.

We introduce the following relationship between mean and variance in the model:

$$(3) \qquad var(y_i|x_i) = \phi\mu_i$$

where ϕ equals the scale parameter. We run the Poisson regression with the following results.

Table 3.1.2 Parameters for Poisson Regression on Number of Complementary Suppressions for a Primary Suppression, σ^2 = 1.851504μ

Parameter	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	-0.8638	0.1820	22.52	<.0001
prot	0.0001	0.0000	231.13	<.0001
$prot^2$	-0.0000	0.0000	137.29	<.0001
numrel	0.7309	0.0833	76.90	<.0001
lnCapsib	-0.1670	0.0097	294.35	<.0001
depth	-0.2000	0.0132	228.49	<.0001
numcomp1	0.0772	0.0030	659.28	<.0001
lnVal	0.2804	0.0115	594.21	<.0001
Numtot	-0.5792	0.0934	38.48	<.0001
Scale	1.3607	0.0000	-	-

The estimate for scale equals the square root of the scale parameter ($\sqrt{\phi} = 1.3607$). Hence we have the following empirical mean-variance relationship.

$$(4) \qquad var(y_i|x) = 1.851504\mu_i.$$

Table 3.1.3 Deviance and Pearson Residuals for Number of Complementary Suppressions for a Primary Supression, σ^2 = 1.851504μ

Criterion	DF	Value	Value/DF
Deviance	5462	10112.9439	1.8515
Scaled Deviance	5462	5462.0000	1.0000
Pearson Chi-Square	5462	11273.6239	2.0640
Scaled Pearson X^2	5462	6088.8831	1.1148

The total deviance of the model is 11984.1585. To compute the R^2 value, Cameron and Windmeijer (1996, 214) propose the following:

(5) $$R^2 = 1 - D(\widehat{\mu}_s, \widehat{\mu}, \widehat{\phi}) / D(\widehat{\mu}_{si}, \widehat{\mu}_0, \widehat{\phi}).$$

The numerator in the second right-hand-side term is the total deviance, the denominator the deviance for the intercept-only model. The intercept-only deviance is available from the GENMOD procedure's type1 option.

Table 3.1.4 Total Deviance, Intercept-only and with Predictors Successively Added, σ^2 = 1.851504μ

Parameter	Deviance
Intercept	28791.5281
prot	19507.3058
$prot^2$	16890.2923
numrel	14955.4817
lnCapsib	14674.6060
depth	13530.3180
numcomp1	11307.8149
lnVal	10190.4357
Numtot	10112.9439

The final model is:

(6) $$\boldsymbol{ln(NUMC) = -.8638 + 0.0001PROT + 0.000PROT^2 + 0.7309NUMREL - 0.167ln(CAPSIB_{AVG}) - 0.2DEPTH + 0.0772NUMCOMP1 + 0.2804ln(VAL) - 0.5792NUMTOT + \varepsilon}$$

All coefficients were statistically significant at the α = 0.001 level.

(7) $$R^2 = 1 - \frac{10112.9439}{28791.5281} = 0.649$$

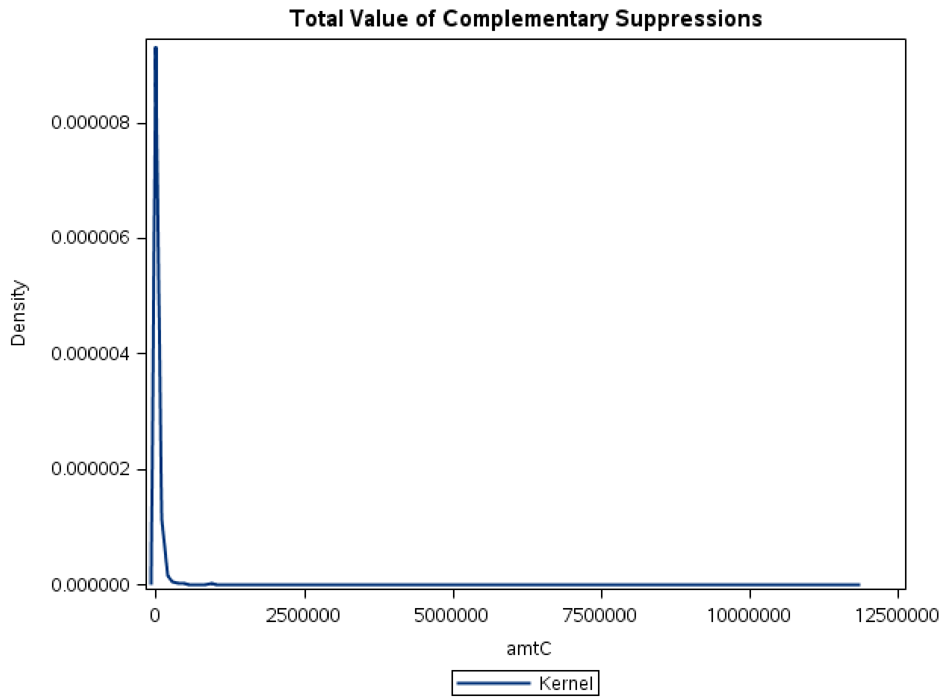
The model explains that 64.9 percent of the variance in the number of complementary cells is suppressed. A Pseudo-R^2, computed by the squared correlation between the fitted and actual values for numC, is:

(8) $$R^2 = \left(corr(\hat{\mu}, y)\right)^2 = 0.842^2 = 0.709.$$

3.1.3 Model for Sum of Complementary Suppressions for a Primary Suppression - Methodology

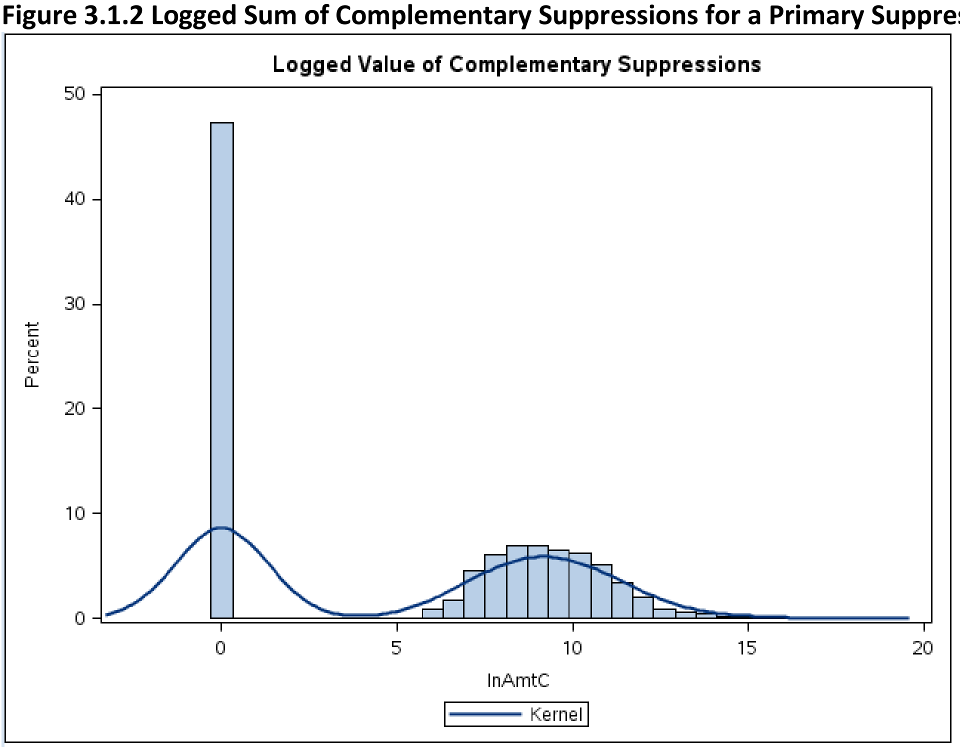
Also of interest is predicting the total number suppressed in complementary suppressions generated by primary suppressions. Compared to the number of complementary suppressions, the amount suppressed has proven more difficult to predict, due to the skewed nature of the data. The data is displayed in the following density plot.

Figure 3.1.1 Total Value (Sum) of Complementary Suppressions for a Primary Suppression



A closer inspection, made by logging the value of complementary suppression, reveals a bimodal distribution.

Figure 3.1.2 Logged Sum of Complementary Suppressions for a Primary Suppression



To obtain the best fit for a predictive equation, we experimented with several transformations of the data (log, squared, and square root) and with a quantile regression, and we determined that an OLS with the square root of the dependent variable delivered the best fit. This approach is also supported by the literature. Vellemann and Hoaglin (1981) suggest lower powers for right-skewed data and higher powers for left-skewed data. Baker (1930) lists three ways to transform a bimodal distribution into a unimodal one, among which is,

$$(9)\quad u = x^n,$$

where n represents the number number of modes. In our case, this translates to:

$$(10)\quad u = x^2,$$

with u as the dependent variable and x representing predictors. To fit the regression we take the root of both sides to yield:

$$(11)\quad \sqrt{u} = x.$$

We propose fitting the following OLS model:

$$\textbf{(12)}\quad \sqrt{\boldsymbol{AMTC}} = \boldsymbol{\beta_o} + \boldsymbol{\beta_1 PROT} + \boldsymbol{\beta_2 PROT^2} + \boldsymbol{\beta_3 NUMREL} + \boldsymbol{\beta_4 ln(CAPSIB_{AVG})} + \boldsymbol{\beta_5 DEPTH} + \boldsymbol{\beta_6 NUMCOMP1} + \boldsymbol{\beta_7 ln(VAL)} + \boldsymbol{\beta_8 NUMTOT} + \boldsymbol{\varepsilon}$$

AMTC – Amount in Complementary Suppressions. The total value of all the complementary suppressions induced by the primary suppression.

3.1.4 Model for Sum of Complementary Suppressions for a Primary Suppression - Evaluation and Results

The model yields the following parameter estimates:

Table 3.1.5 Paremeters for OLS for Sum of Complementary Suppressions

Parameter	Parameter Estimate	Standard Error	t-value	Pr > |t|
Intercept	-243.69160	32.53854	-7.49	<.0001
prot	0.03708	0.00215	17.21	<.0001
$prot^2$	$-2.3774E^{-7}$	$4.490626E^{-8}$	-5.27	<.0001
numrel	96.43897	15.38463	6.27	<.0001
lnCapsib	-9.19193	1.26705	-7.25	<.0001
depth	7.72007	1.87857	4.11	<.0001
numcomp1	14.84923	0.74601	19.90	<.0001
lnVal	6.57230	1.46232	4.49	<.0001
Numtot	-32.52415	16.30120	-2.00	<.0461

and the related fit statistics:

Table 3.1.6 Fit Statistics for OLS for Sum of Complementary Suppressions

Statistic	Value
Root MSE	153.85311
Dependent Mean	86.36742
R-Squared	0.3676
Adj R-Squared	0.3666

Hence we arrive at the following model:

$$\textbf{(13)}\quad \sqrt{amtC} = -243.6916 + .03708PROT - 2.377\times 10^{-7}PROT^{2} + 96.43897NUMREL - 9.19193\ln(CAPSIB_{AVG}) + 7.72007DEPTH + 14.84923NUMCOMP1 + 6.5723\ln(VAL) - 32.52415NUMTOT + \varepsilon$$

All variables were statistically significant ($p<0.0001$ except for NUMTOT, where $p<0.05$). The positive coefficient (7.72007) on *DEPTH* is consistent with findings such as Wang (2016), who showed that depth matters.

3.2 Empirical Study

3.2.1 Methodology for Empirical Study

The current production program for cell suppression uses sequential estimation of complementary suppressions given a list of primary suppressions. The order of primary suppressions (primaries) sent to the program affects the number and order of complementary suppressions selected.

The Cell Suppression group at the Census Bureau suggested running a series of cell suppressions on a "small" portion of the 2012 Economic Census data and calculating the average number and sum of secondary suppressions. We used the additive Chernoff bound (Dwork et al, 2013, page 29) to determine the number of data suppression runs to estimate these averages, and later to determine the probability that each empirical average was within a neighborhood of the population average.

Theorem (Additive Chernoff Bound [Dwork et al, 2013, page 29]). Let $X_1, \dots, X_m$ be independent random variables bounded such that $0 \le X_i \le 1$ for each i. Let $S = \frac{\sum_{i=1}^{m} X_i}{m}$ denote their mean and $\mu = E[S]$ be their expected mean. Then:

$$Pr[S > \mu + \varepsilon] \le e^{-2m\varepsilon^2}$$
$$Pr[S < \mu - \varepsilon] \le e^{-2m\varepsilon^2}$$

To apply this theorem to the number of complementary suppressions, let X_i be the ratio of the number of complementary suppressions found to the number of cells. For the sum of complementary suppressions, let X_i be the ratio of the sum of complementary suppressions to the sum of all cell values.

Table 3.2.1 provides the value of ε to be 90-percent certain that an average is within ε of the population average given number of random runs *m*.

Table 3.2.1 Additive Chernoff Bounds for 90 Percent p

p	m	$\varepsilon = \sqrt{\frac{-ln(p)}{2m}}$
0.90	20	0.0513
0.90	30	0.0419
0.90	40	0.0363
0.90	50	0.0325

Given time and resource constraints, we decided to do only 50 random runs of the cell suppression program. From table 3.2.1 we would be 90-percent confident that our estimates would be within 3.25 percent of the population averages. We needed seven files were to conduct each experiment: the parameter file, the input file, the file of primaries, three relationship files, and a dummy file. Each experiment was expected to take around six hours.

We conducted both random and non-random runs. The orders for the non-random runs were based on protection or depth by protection. To explore further the sensitivity of results to ordering, we ran additional runs for permutations of the top three cells for three of the non-random runs.

3.2.2 Findings from Empirical Study

We used annual payroll data for the Plastics and Rubber Manufacturing, subsector 326 (NAICS codes starting with the digits 326), from the 2012 Economic Census.

Due to time constraints, the experiment stopped at 50 random runs. Solving for ε given m (50) and p (0.90), one is 90-percent confident that each average is within 3.25 percent of the associated population average.

The empirical average number of complementary suppressions for the 50 cell suppression runs was 14155.44, with a standard deviation of 1.05. The empirical average sum of complementary suppressions for the 50-run cell suppression experiment was 266,687,403.70, with a standard deviation of 45,965.58. The run times were between 5 and 6.3 hours. Figures 3.2.1 and 3.2.2 present the running average count and sum of complementary suppressions, respectively, for the 50 random runs.

Figure 3.2.1

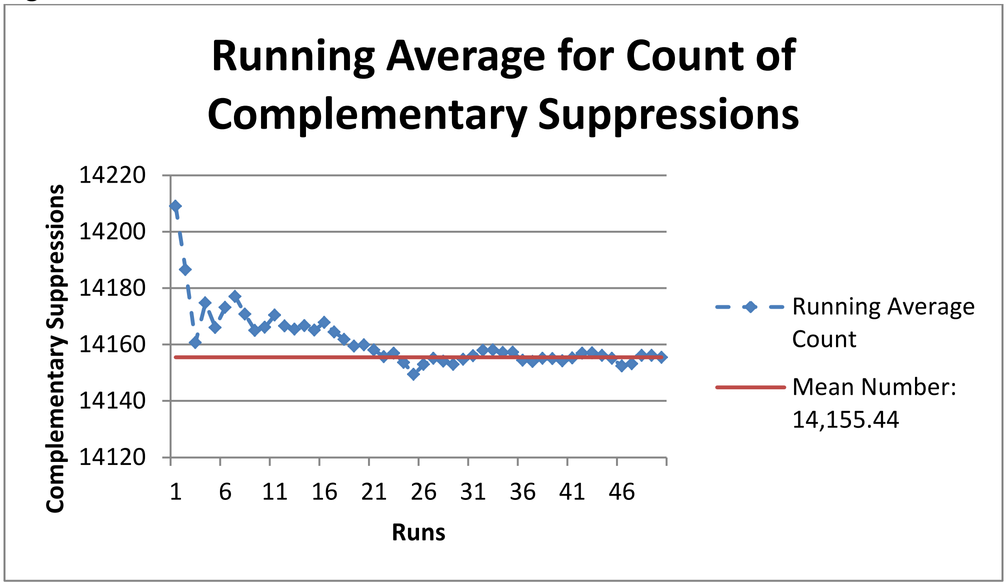


Figure 3.2.2

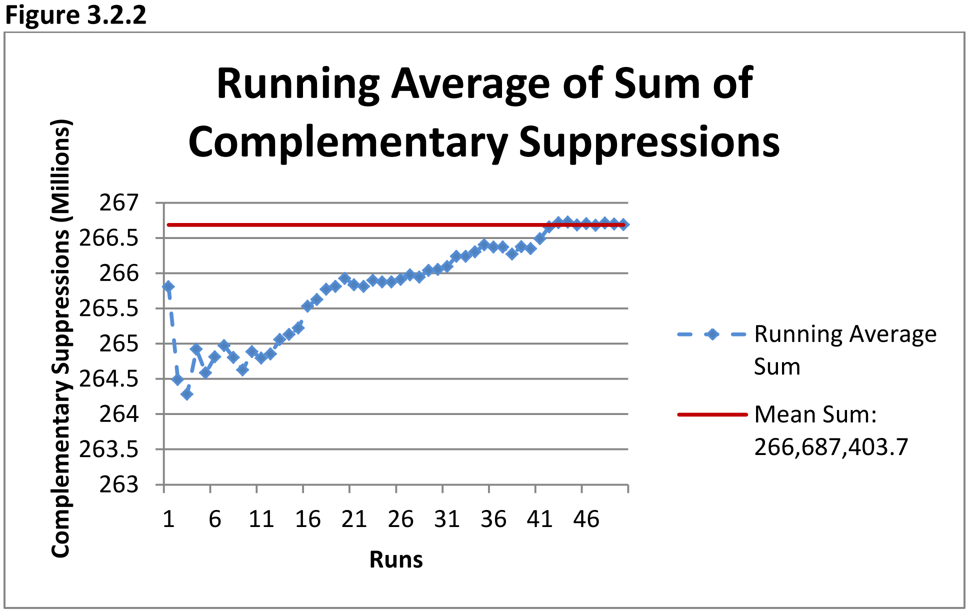


Table 3.2.2 presents the number and sum of unduplicated complementary suppressions for the non-random runs of subsector 326, as well as the mean, maximum, and minimum statistics for the 50 random runs. The rows for the non-random runs are for a single run each. The maximum number and sum of complementary suppressions on the random runs were achieved on different runs, as were the minimum number and sum of complementary suppressions.

Table 3.2.2 Number and Sum of Unduplicated Complementary Suppressions for Non-Random Runs, Average Number and Sum for Random Runs

Sort Order	Number of Unduplicated Cs	Sum of Unduplicated Cs
Non-Random Runs		
Increasing protection	13938	260489750
Increasing depth, Decreasing protection	14010	268638135
Increasing protection, Decreasing depth	14031	264328786
Decreasing protection, Increasing depth	14122	264693203
Decreasing depth, Increasing protection	14197	262967776
Decreasing protection	14589	274997405
Random Runs		
Mean	14155.44	266687403.70
Maximum	14295	273536471
Minimum	14030	262498160

Figures 3.2.3 and 3.2.4 present the number and sum of complementary suppressions (Cs), respectively, for non-random runs. Within a given ordering, the runs are ordered by permutations of the top three cells. The figures also present the maximum and minimum number and sum of Cs of the 50 random runs. It appears from Figures 3.2.3 and 3.2.4 that reordering the top three cells affects the number and sum of Cs.

Figure 3.2.3

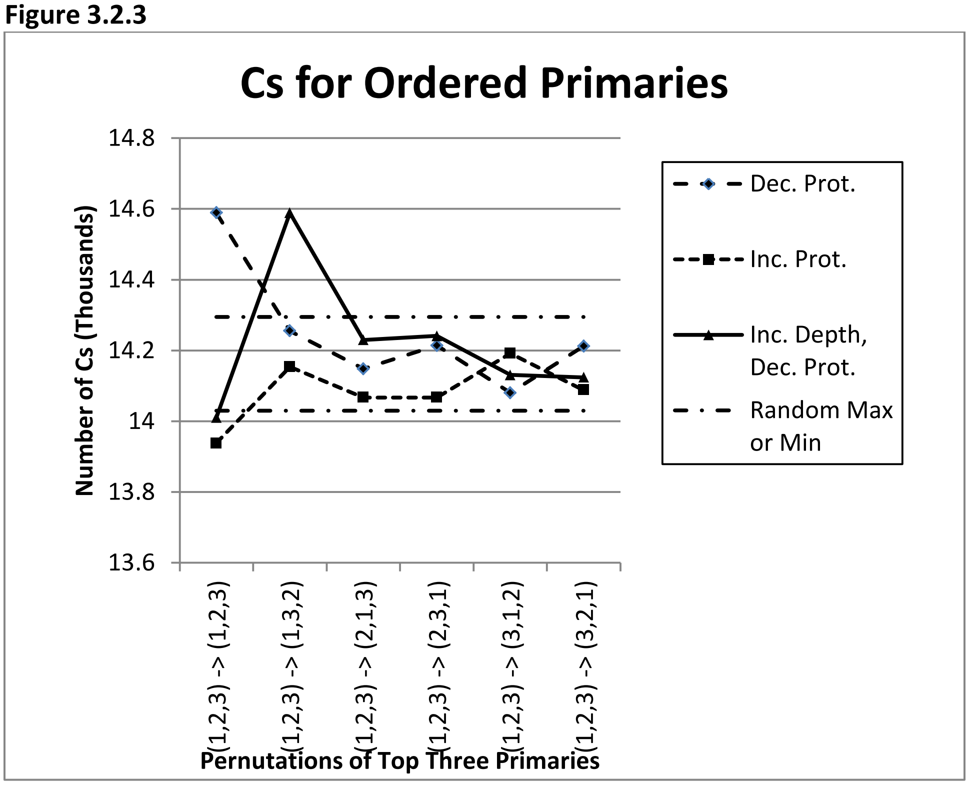
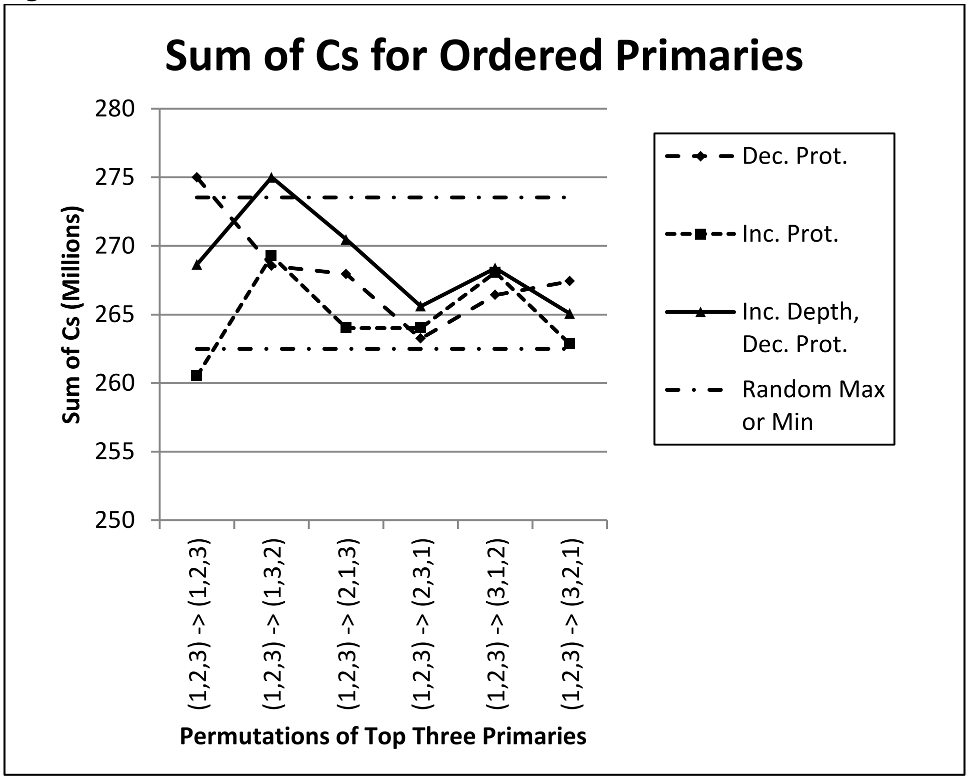


Figure 3.3.4



4 Conclusions and Next Steps

Model results confirmed prior findings regarding depth and protection value for sector 72. As expected, conducting the empirical study for subsector 326 used many resources.

Further research could help answer whether the models developed in this research apply to other sectors and other surveys. Applying the models to sector 326 would be a step towards answering that question. The models produced for this paper were based on 2-dimensional data (NAICS, geography) from sector 72, but the empirical results were based on 3-dimension data (NAICS, geography, contents) from subsector 326. A positive result, with no changes required for the models, would be quite promising. If changes are required in the models, then one might explore whether the changes were due to an increase in dimensionality. Either result would be useful for planning.

5 References

Baker, G. A. 1930. “Transformations of bimodal distributions.” *Annals of Mathematical Statistics* 1: 334-344. Available at https://www.jstor.org/stable/pdf/2957637.pdf (accessed August 2016).

Cameron, A. C. and F. A. G. Windmeijer. 1996. “R-squared measures for count data regression models with applications to health-care utilization.,” *Journal of Business & Economic* Statistics 14: 209-220. Available at doi: 10.2307/1392433 (accessed July 2016).

Cox, L., J. T. Fagan, B. V. Greenberg, and R. J. Hemmig. 1986. “Research at the Census Bureau into Disclosure Avoidance Techniques for Tabular Data.” Paper presented at the annual meeting of the American Statistical Association Survey Research Methods Division, Chicago, Illinois, August 18-21.

Cox, L. 1975. “Disclosure Analysis and Cell Suppression.” Paper presented at the annual meeting of the American Statistical Association Social Statistics Division, Atlanta, Georgia, August 25-28.

Dwork, C. and A. Roth. 2014. “The Algorithmic Foundations of Differential Privacy.” Foundations and Trends in Theoretical Computer Science 9: 1–277. Available at http://isites.harvard.edu/fs/docs/icb.topic1465468.files/The%20Algorithmic%20Foundations%20of%20Differential%20Privacy.pdf (accessed August 2016).

Federal Committee on Statistical Methodology. 2005. *Statistical Policy Working Paper 22: Report on Statistical Disclosure Limitation Methodology.* Statistical and Science Policy, Office of Information and Regulatory Affairs, Office of Management and Budget, Washington, DC. Available at https://www.hhs.gov/sites/default/files/spwp22.pdf (accessed September 2016).

Massell, P. 2002. “Optimization Models and Programs for Cell Suppression in Statistical Tables.” Presentation at the Annual Meeting of the American Statistical Association, New York, New York, August 10-13, 2002.

Steel, Phil, J. Fagan, V. Harusadangkul, P. Massell, R. Moore Jr., J. Slanta, and B. Wang. 2013. “Techniques to apply cell suppression to large sparse linked tables and some results using those techniques on the 2013 (US) Economic Census.” United Nations Economic Commission for Europe Working Paper.” Available at https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2013/Topic_1_USA.pdf (accessed August 2016).

Velleman, P. F. and D. C. Hoaglin. 1981. Applications, basics, and computing of exploratory data analysis. Pacific Grove, CA: Duxbury Press.

Wang, Bei. 2013. “Improving LP performance in cell suppression process.” Presentation at the Joint Statistical Meetings of the American Statistical Association, Montreal, Quebec, Canada, August 4, 2013.

———. 2016. “An ordering that improves the cell suppression pattern in sequential processing.” Presentation at the Joint Statistical Meetings of the American Statistical Association, Chicago, IL, July 31, 2016.