Working Paper
ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Ottawa, Canada, 28-30 October 2013)

Topic (iv): The trade-off between quality, utility and privacy

Statistical Analysis of Suppressed Tabular Data

Prepared by Lawrence H. Cox, National Institute of Statistical Sciences, United States of America (cox@niss.org)

Statistical analysis of suppressed tabular data

Lawrence H. Cox*

*National Institute of Statistical Sciences, Durham, NC 27709, USA, cox@niss.org

Abstract: Theoretical methods and software are available for performing optimal complementary cell suppression (CCS) in tables. The released resulting suppression patterns comprise algebraic circuits which define alternative tables for the original table while controlling variation between original and alternative cell values. For an important class of statistical tables including two-way tables, these circuits are simple alternating (+/-) cycles. Suppressed tables are notoriously difficult to analyze statistically. A user with sufficient resources could construct the full set, a subset or a probability sample of alternative tables and analyze these tables, resulting in bounds for or estimates of analytical outcomes for the original table. We explore the use of circuits for analysis of suppressed tabular data.

1 Introduction

Since the 1940s, National Statistical Offices (NSO) have employed various forms of cell suppression for disclosure limitation of *tabular data*--data organized in additive tables. Dealing with suppressions in tabular data is difficult, particularly for the less sophisticated analyst. Statistical organizations suppress tabular cells to flag missing or unreliable data, or for confidentiality purposes. How to use or deconstruct suppressed data or tables and assign suitable surrogate values to suppressions so that appropriate statistical analysis can be performed is a difficult problem with underpinnings in Bayesian modeling and algebraic statistics. Whether to construct a single surrogate for analysis or select a sample of surrogates and combine individual analyses—and how—is a central technical question. Doing so in a statistically principled manner—particularly for tables of establishment-type magnitude data—and accounting for uncertainties and bias due to suppression and assignment and combination of surrogates is an unsolved problem. A companion problem is how to measure and evaluate the conformance of estimates and inferences based on surrogates with those based on original data. These topics form the basis of our interests.

For frequency tabulations, two classical indirect approaches to table reconstruction are to invoke iterative proportional fitting or fit a log-linear model and use the MLE as a surrogate for the original table. We discuss a third and the first direct approach: to construct a set of *alternative tables* based on algebraic *moves* from the original table, together with associated probabilities. Analysis then can be performed on a surrogate table or a probability sample of tables and combined.

For magnitude data, such as establishment-based economic census or survey data, the terrain is far less studied. Most users of magnitude data "analyze" tabular magnitude

data one cell value or aggregation at a time, e.g., analyst interest may be confined to cell values within one industry. If these values are suppressed, analysis is thwarted.

Our principal focus is to evolve statistically principled analytical methods leading to a path for inference for tabular data subject to suppression—establishment-type magnitude data in particular. The goal is to develop and provide methods of table deconstruction, analysis, and combination of analytical outputs that are practical and suitable for use in realistic tables by a broad class of users, and to provide methods that statistical agencies can use to construct suitable alternative tables for release in lieu of tables with suppressions. This research has strong *transparency* implications regarding what agencies should (or should not) tell users about disclosure rules and suppression methods and what constitutes *safe release* of a set of alternative tables.

The intruder's task is to deconstruct suppressed entries using prior information, domain knowledge, and information on disclosure and suppression rules and data quality criteria. The analyst, too, can benefit from deconstructing suppressed data, leading to a set of alternative tables. The analyst then performs the analysis on a selected or constructed surrogate table, or selects all or a probability sample of alternative tables, analyzes each, and combines analyses and estimates of uncertainty to produce inferences conformal with those from analysis of the original (true) table.

The concept of *table deconstruction* introduced here includes:

- identifying or estimating feasible values of suppressed entries
- identifying alternative feasible tables to the original table
- ruling out otherwise feasible values or tables based on prior information, deterministic analysis, or probabilistic analysis
- identifying alternative tables expected to be exchangeable with the original table for inferential purposes

2 Complementary cell suppression: Background

Tabular data are data exhibiting an additive structure between subsets of data items, referred to as the tabular cells. Tabular structure can be expressed as a system of linear equations $\mathbf{Ax} = \mathbf{b}$. The constant right-hand side satisfies $\mathbf{b} \geq 0$ and entries of the coefficient matrix $\mathbf{A}$ are restricted to {-1, 0, +1} with each row containing at most one -1. Each row equation describes an aggregation of cell values (coefficients = +1) to a *marginal total* value (coefficient = -1) or to a constant. In official statistics, typically one is concerned with solutions $\mathbf{x} \geq \mathbf{0}$, and often only fully integer solutions

(e.g., contingency tables). Data pertaining to individual entities such as persons or businesses are collected and aggregated in various ways for presentation in tabular form. Person data are often—but not always—count data presented as contingency tables for which each individual contributes 1 to a cell value if its characteristics match those defining the cell, and 0 otherwise. Establishment data often present an aggregate of a nonnegative quantity of interest—such as gross monthly retail sales—over all establishments meeting the criteria defining the cell. These aggregates are referred to as *magnitude data*. Contingency tables based on establishment counts or magnitude data based on individuals' aggregates are also commonplace.

Aggregation systems can be as simple as individual one- or two-way contingency tables, or as large, but well-structured, as national censuses, or as complex as tables from multiple sources linked through aggregation at different levels. Aggregation corresponds to the constraint system of a suitable linear program (continuous data) or mixed integer linear program (integer data). Various problems in official statistics, particularly confidentiality problems, begin with an original table **a** that must be amended (to **a'**) prior to public release. Amendment amounts to modifying all or some original values or replacing certain original values by variables. The latter approach is *cell suppression*, the focus of the proposed research. The amended table must also satisfy the same tabular constraints (viz., **Aa** = **Aa'**). Such problems are decision problems, and consequently for both integer (count) and noninteger (magnitude) data the resulting computational problem is a mixed (binary) integer linear program likely to be intractable to solve to optimality in general.

Release of any statistical information based on data pertaining to individual subjects poses some *risk of disclosure* of individual subject data. It is the responsibility of the statistical office to distinguish between acceptable and unacceptable disclosure (the *disclosure rule*) and employ effective disclosure limitation methods (the *suppression rule* or algorithm) to address the latter. Release of the *true* value of a tabular cell may pose an unacceptable risk of disclosure if the cell covers too few respondents or its value is highly *dominated* by a small number of respondents.

As far back as the 1940s, NSOs have employed various notions of cell suppression for disclosure limitation of tabular data. The process of c*omplementary cell suppression* (**CCS**) may be summarized as follows. First, cells representing unacceptable risk of disclosure (disclosure cells, *primary disclosures*, or *sensitive cells*) are identified and their risk quantified by a disclosure rule. Based on the numeric value of this risk, safe bounds $0 \leq l_x \leq x$ and $u_x \geq x$ protecting the value x of each sensitive cell X are identified. The *risk region* is the open interval (l_x, u_x). Second, disclosure cells are removed (suppressed) from publication. This is primary cell suppression. Third, owing to aggregation relationships among cell values, it is

unlikely that primary suppression alone will reduce disclosure risk sufficiently to allow release of the tabular data. Thus, additional, nonsensitive cells must be suppressed to achieve a sufficient disclosure limitation. Suppression has the effect of replacing constant values with variables in the linear system $\mathbf{Ax} = \mathbf{b}$. An alternative table is a nonnegative (often, integer) solution of the new system. The sacrificed cells are the *complementary suppressions*. Optimal selection of complementary suppressions is an NP-hard problem, even for one-dimensional tables.

Sufficient protection is defined in terms of a set of alternative tables satisfying the following conditions. For each sensitive cell **X**, select l_x (respectively, u_x). If there does not exists an alternative table exhibiting a value x for **X** satisfying $x \leq l_x$ (respectively, $x \geq u_x$), protection is not sufficient. In other words, sufficient protection amounts to ensuring that among the alternative tables, for each sensitive cell **X,** values to the left and to the right of its risk region (l_x, u_x) are exhibited.

The state-of-practice of complementary cell suppression (**CCS**) is broad and uneven. At one extreme, only primary suppression is performed. This remains the case for a variety and number of state data and certain federal data in the U.S., notably public health data. Or, subject matter experts may perform complementary suppression *by-hand*. Towards the center, software which operationalizes by-hand reasoning is employed. Often, this amounts to little more than enhancing the speed and effects of flawed approaches, viz., doing the wrong thing faster. Finally, principled methods or sound heuristics based on mathematical and statistical science are invoked. The latter is the case at many, but by no means all NSOs, with considerable variability in the characteristics and effectiveness of the methodology and software. Despite compelling evidence against using CCS on both confidentiality and usability grounds, overwhelmingly cell suppression continues to be used.

Complementary cell suppression performed using modern methods and software is consistent internally, meaning that it provides sufficient protection. Consistency relies upon weighted *circuits* between suppressed entries that enable movement from one alternative table to another (viz., between nonnegative solutions of $\mathbf{Ax} = \mathbf{b}$). However, CCS is vulnerable to intruder attack based on applying outside intruder knowledge to these circuits.

Expressed within a mathematical programming framework, complementary cell suppression is driven by data loss criteria expressed as constraints on individual cell values that preserve *local data quality* and a linear cost function based on a suitable notion of *global data quality*. The objective of CCS is to identify a suppression pattern that, subject to fulfilling data protection requirements, represents a solution of maximal or near-maximal data quality. However, even "optimal" suppression solutions replace some original values by "holes" in the data and thereby thwart data utility and analysis, particularly for the less sophisticated user. Also, even

sophisticated users are on shaky ground, as data missing-ness is due to mathematical relationships and choices not based on any (discernible) probability model for missing-ness. It is not clear how users cope with these problems when analyzing suppressed tables, or what improved strategies might be. This motivates the principal issue to be addressed by the proposed research—methods for statistical analysis and a pathway to inference on tabular data subject to suppressions, including analysis of uncertainty and bias stemming from suppression and imputation.

3 Mathematical basis for CCS: Circuits

Disclosure limitation via CCS is based on *circuits* between sets of suppressed cells. In the simplest cases (tables of network type including two-way tables7), a circuit admits a bounded (network) flow between its constituent suppressed entries. Consider the following example:

D_{11} (1)	18	D_{13} (6)	**25**
13	D_{22} (5)	**D_{23} (2)**	**20**
D_{31} (4)	**D_{32} (1)**	10	**15**
18	**24**	**18**	**60**

Table 1: Table with Suppressions Containing 4 Sensitive Cells

The original (true) values of suppressed cells appear in parentheses. For concreteness, assume disclosure is defined by (integer) cell values in the risk region $(0, 5) = \{1, 2, 3, 4\}$. This is known as the 5-threshold disclosure rule, a specific case of the *t-threshold rule*, $t > 0$. Table 1 contains 4 disclosure cells (bolded) and 2 complementary cells. This suppression pattern is optimal with respect to both minimizing number of suppressions (6) and minimizing total value of suppressions (19). As a mathematical problem, this suppression pattern is equivalent to:

D_{11} (1)	0	D_{13} (6)	**7**
0	D_{22} (5)	**D_{23} (2)**	**7**
D_{31} (4)	**D_{32} (1)**	0	**5**
5	**6**	**8**	**19**

Table 1a: Reduction of Table 1

The pattern corresponds to a single circuit of the form

+/-	0	-/+
0	-/+	+/-
-/+	+/-	0

Table 1b: Circuit for Table 1 Suppressions

The circuit is interpreted as follows. Relative to D_{11}, up to four units can be moved in the + direction along this circuit (increasing the value of D_{11}) and up to one unit in the – direction, yielding 6 alternative values for D_{11} and 6 alternative integer tables (including the original table), corresponding to D_{11} = 0, 1, …, 5. If d = -1, 0, …, 4 units are moved through D_{11}**,** then d units are also moved through D_{23} and D_{32} while $-d$ units are moved through D_{13}, D_{22} and D_{31}**.** The 3 cells in each of the two respective sets of cells are said to be of the *same parity*. Among the 6 alternative tables, each sensitive cell achieves a value outside the risk region (1, 4), so sufficient protection is provided by this suppression pattern.

For higher-dimensional tables, circuits can be more complex than simple +/- paths, and involves algebraic statistics. For tables regarded as realistic from the standpoint of censuses and surveys, these paths will be investigated as part of the table deconstruction research. Many realistic tables are two-way tables or tables that are similar mathematically to two-way tables (*tables of network type*), 2-way tables

- hierarchies of 2-way tables organized along one, but not both, dimensions. Important examples are tables for an industry group organized in a geographic hierarchy, and tables for a geographic region organized hierarchically by all industries
- "thin" multi-way tables: tables of size bxcx2x2x….x2; e.g., a 4-way table comprising 5 racial groups x 9 income categories x 2 genders x 2 geographic (urban/rural)

4 Alternative tables

Table 1 was shown to offer 6 alternative tables of nonnegative counts consistent with the suppression pattern, corresponding to $D_{11} = 0, 1, \ldots, 5$. The true (original) table has $D_{11} = 1$.

What is and what is not a sensitive cell (primary disclosure) differs between these tables as the value of D_{11} changes. Consider the 5 other alternative tables, with notation primary (**P**) and complementary (**C**). First, the $D_{11} = 0$ table:

Alternative table **(**$D_{11} = 0$)

0	18	7	25
13	6	1(P)	20
5	0	10	15
18	24	18	60

Optimal suppression pattern

0	18	7	25
13	**D**	**D**	20
5	**D**	**D**	15
18	24	18	60

or

0	18	7	25
D	6	**D**	20
D	0	**D**	15
18	24	18	60

The $D_{11} = 0$ table contains only 1 sensitive cell (P) under the 5-threshold rule. Two alternative suppression patterns illustrate cases where cells of value 0 are allowed as complementary suppressions, and when they are not. The first pattern is optimal with respect to both number of suppressions and total value of suppressions. The second is also optimal for both criteria, after exempting 0-cells from suppression. Both tables exhibit a pattern different from the released pattern, and thus the $D_{11} = 0$ table could <u>not</u> have been the original table, thereby reducing from 6 to 5 the number of tables that could have been the original table. Next, consider the $D_{11} = 2$ table:

Alternative table ($D_{11} = 2$)

2**(**P)	18	5	25
13	4(P)	3(P)	20
3(P)	2(P)	10	15
18	24	18	60

Optimal suppression pattern

D	18	D	25
13	**D**	**D**	20
D	**D**	**10**	15
18	24	18	60

This table contains 5 primary and 1 complementary suppressions, and the resulting pattern is identical to the original pattern. So, $D_{11} = 2$ is not yet ruled out. However, if the agency were to distinguish between primaries and complementaries—or if by some means the intruder could do so--then the intruder/analyst would detect the difference, and could rule out the $D_{11} = 2$ alternative table.

Original Optimal Pattern ($D_{11} = 1$)

P	18	C	25
13	C	P	20
P	P	10	15
18	24	18	60

Optimal Pattern ($D_{11} = 2$)

P	18	C	25
13	**P**	**P**	20
P	**P**	**10**	15
18	24	18	60

Indeed, the same can be said of the remaining three alternative tables: $D_{11} = 3, 4, 5$. Thus, if primaries are or can be identified, all tables other than the original (true) table have been ruled out as alternatives, with the result that **no disclosure protection** has been achieved. This discussion illustrates the complexity of *transparency issues*—issues surrounding what, if any, ancillary information the agency provides regarding its disclosure and/or suppression rules.

In general, statistical agencies **do not** distinguish between primary and complementary suppressions in released tables. The preceding examples demonstrate **why** this is an important and correct approach based on confidentiality—and may in fact be new. The problem remains that, based on domain or outside knowledge or other information, it may be possible for the intruder/analyst to identify some primaries—or complements—and thereby rule out one or more alternative tables. In particular, for economic magnitude tabulations such as from economic censuses, it is commonplace for the agency to release the establishment count for all cells, including primary disclosures. Thus, cells with establishment count 1 or 2 are de facto primaries and thereby identified.
This discussion raises three transparency issues. The first, answered in the negative by our analysis, is whether primaries should be identified. The second is whether the agency should release the disclosure rule, viz., the value t = 5 in the illustration or the value of p in a p-percent rule. The third is whether the agency should release details on how it selects complementary suppressions, viz., minimum number of suppressions or minimum total value of suppressions in the illustration.

5 Table deconstruction and analysis of suppressed tables

The preceding section demonstrates how complementary suppression can be undone by "too much" transparency. The reasoning was deterministic. It is also possible to rule out alternative tables through statistical or probabilistic analysis and inferential considerations. For an example, multiply all entries of Table 1 by 100 and treat it as a table of suppressed magnitude data subject to an unspecified disclosure rule.

With these modifications, the conditional chi-square statistic is appropriate to compare an alternative set of table values $\{c_i\}$ with the original (true) values $\{a_i\}$:

$$\chi^2_{(df)} = \sum_i \frac{(c_i - a_i)^2}{a_i}$$ *df* is the number of degrees of freedom

For a table subject to suppression, the indices *i* are restricted to the suppressed entries. If the suppression pattern consists of a single circuit, as in Table 1, then df = 1, corresponding to an integer quantity *d* that can be moved around the circuit without violating nonnegativity. Thus, $c_i = a_i \pm d$. For modified Table 1 (D_{11} = 100), $-100 \leq d \leq 400$. Thus, relative to any alternative table $\{a_i\}$:

$$\chi^2_{(1)} = \sum_i \frac{d^2}{a_i} = d^2 \sum_i \frac{1}{a_i}$$

For modified Table 1, the sum of reciprocals of suppressed true entries is 0.032, and for $\alpha = 0.05$ the critical chi-square value is 3.84. For $d^2 > 120$, the chi-square statistic exceeds the critical value. Thus, alternative tables with $|d| \geq 12$ are not reliable substitutes for modified Table 1 (D_{11} = 100) and are ruled out on inferential grounds as surrogates or as contributors to a constructed surrogate. Similarly, any scheme for combining analyses of alternative tables must look skeptically at these tables. Thus, (only) on the basis of preserving analytical and inferential outcomes, it would appear that choices for a set of analytical substitutes to modified Table 1 or for combining analyses are limited to $88 < D_{11} < 112$. For integer tables, there are 23 suitable choices among 501 alternative tables, viz., $D_{11} = 0, \ldots, 500$.

These observations are of crucial importance: it is <u>not</u> necessary to construct all alternative tables but only a sufficiently large subset or probability sample of tables expected to be exchangeable with the original table for inferential purposes. This should drastically reduce the computational burden of any algebraic methods that need to be applied, e.g., from 501 to 23 tables.

The preceding analysis is from the perspective of the statistical agency which knows that D_{11} = 100 corresponds to the true modified Table 1. The agency may or may not choose to provide analysts with information on selection of suitable alternative tables (a transparency issue). Doing so, the agency should avoid information that is symmetric about the true value.

The analyst does not know which table is the original table among the 501 possible choices. From the analyst's perspective, more information is needed. The analyst's information needs are analogous to those of an intruder seeking to exploit vulnerabilities of CCS. Below are sketched some approaches for the analyst. The proposed research will include full examination of these and other approaches.

The analyst can identify the 501 alternative tables to Table 1. If, based on domain knowledge, the analyst can derive effective lower and/or upper bounds on any suppressed cell value(s), then the analyst can reduce the range of D_{11}. For magnitude data, this is often possible. For example, if the analyst can determine that $D_{11} \geq 70$ and $D_{23} \leq 550$, only 81 choices remain, viz., $70 \leq D_{11} \leq 150$. Knowledge of the disclosure and suppression rules also can be applied. For example, the analyst may infer D_{11} to within some percentage (5,%,10%, 15%, …) of the largest (primary) cell value because, based on a p-percent rule, a primary cell value is replaced by a safe value at a distance of at most p-percent of its original value. Note that revealing or approximating any value along a circuit reveals or provides a corresponding estimate of any other value along the circuit.

The analyst could compute all 501 sum of reciprocal values and the corresponding chi-square statistics, from which clusters of similar alternative tables could be identified and ranges for D_{11} narrowed. Or, (s)he could examine the behavior of the sum of reciprocals of suppressed entries for any alternative table in the current range and possibly develop bounds on this quantity to compare potential chi-square values with the critical value, leading to a "generic" *d* and a small range for it, as follows.

The sums of squares of reciprocal values are of the form: $\sum_i \frac{1}{a_i \pm d}$. As

$$\left|\frac{1}{a_i} - \frac{1}{a_i \pm d}\right| = \frac{|d|}{a_i(a_i \pm d)}$$

and as, for magnitude data, often $|d| << a_i$ for most i, then it may be that the sums of squares of reciprocals do not differ greatly from each other. Consequently, the analyst might be able to compute a single range for a generic *d* centered on any alternative table without knowing which alternative table is the original table.

For example, consider the 4 alternative tables with D_{11} = 100, 200, 300, 400. The respective sums of reciprocals are: 0.03117, 0.0211, 0.0199 and 0.0253. A conditional chi-square relative to these four original tables yields ranges for allowable displacements *d* of each table of (-12, 12), (-14, 14), (-14, 14) and (-13, 13), respectively. The analyst might thus adopt the range $-14 < d < 14$ on a generic *d* centered on <u>any</u> alternative table. The analyst next might, by other means, infer one or more clusters of (27) tables containing the original table. Alternatively, the NSO may choose to release a set or sample of alternative tables with inferential properties similar to the original table in lieu of releasing a table with suppressions.