Estimation for Cells Suppressed in Tabulation with Application to Output Disclosure Treatment of the NSF Survey of Earned Doctorates

A.C. Singh, J.M. Borton, S.H. Cohen, V.E. Welch, Jr.,
B. Groenhout, and Y. Lin
[1]NORC at the University of Chicago, Chicago, IL 60603
[2]National Center for Science and Engineering Statistics, NSF, Arlington, VA 22230
singh-avi@norc.org; borton-joshua@norc.org; scohen@nsf.gov; welch-vince@norc.org; groenhout-brianna@norc.org; lin-yongheng@norc.org

Abstract

National Center for Science and Engineering Statistics (NCSES) employs the method of cell suppression in tabulation (or cs-tabulation) for disclosure treatment of small sensitive cells in tables based on data from the Survey of Earned Doctorates (SED). Since in cs-tabulation, the actual total of suppression partner cell counts can be deduced from unsuppressed cells and cell aggregates, it has the desirable property of preserving true totals of all suppressed and unsuppressed disclosure-safe cell and cell aggregate counts. However, there are several concerns with cs-tabulation that need to be addressed. First, for high dimensional tables, it is computationally difficult in general to develop an efficient and systematic system for finding suppression partner cells corresponding to a given threshold such that suppressed cell aggregates remain analytically or substantively meaningful. Second, it is difficult to perform a valid data analysis in the presence of suppressed cells because different users may fill-in different values for such cells from unsuppressed cell and cell aggregate counts using their own ad hoc ways leading possibly to inconsistent conclusions. There is a need for the data producer to fill-in best estimates for suppressed cells using only unsuppressed information so that any negative impact on user's ability to analyze estimates for domains involving suppressed cells such as underrepresented minority of race/ethnicity groups and women is minimized. This process should maintain internal consistency in output from repeated user queries in that estimated values of cells released earlier are preserved if subsequent queries for tabular output contain the same cells. Third, it may be possible to "reverse-engineer" and fill in some suppressed cells with values that are too close to true values because of a large number of algebraic constraints imposed on values of suppressed cells by unsuppressed cells and cell aggregates. By increasing the confidentiality threshold, more cells can be designated as suppressed cells which will make it more difficult to obtain precise estimates of suppressed cells. Clearly, there is also a need for disclosure audit to check adequacy of the threshold which might lead to its revision. We propose a new method termed estimation for cells suppressed in tabulation (ECS-Tab) that addresses the three concerns mentioned above for cs-tabulation. Simple real-sounding hypothetical examples based on SED are used to illustrate the proposed method.

Key Words: cell suppression; estimation for cells suppressed; log-linear modeling; small cell disclosure; tabular output disclosure treatment

1. Introduction

The cell suppression in tabulation (cs-tabulation) method for protecting confidentiality of small cell counts in tabular output is one of the earliest approaches to statistical disclosure control (Cox, 1980, Fischetti and Salazar, 2000) and is currently used by NSF-NCSES for disclosure treatment of small sensitive cells in tables based on

data from SED. Although it has the desirable property of preserving true totals of all suppressed and unsuppressed disclosure-safe cell and cell aggregate counts, there are several concerns that need to be addressed. In this paper, we propose a new method termed estimation of cells suppressed in tabulation (ECS-Tab) to overcome several limitations of cs-tabulation. In the publication of tables from SED by NCSES, there is concern about the breach of the confidentiality pledge to respondents if a user can identify an individual as a respondent, or can learn about an attribute of a respondent to the survey. This could happen if there is a small cell count in a sensitive cell; that is, the cell count is below a prescribed confidentiality threshold such as 5. For example, consider a hypothetical Table 1(a) for post-graduation plans of doctorate recipients in life sciences by race/ethnicity based on SED data. There are several cells that are too small for publication using the threshold of 5. However, some cells with counts less than 5 may be deemed to be insensitive and hence disclosure-safe such as cell counts in other/unknown categories. Thus, there are four cells under the American Indian or Alaska Native (AI/AN) column and two cells under the mixed race column which are disclosure-prone.

To deal with the problem of small sensitive cell counts in any tabular output, the method of cs-tabulation suppresses cells with counts below the threshold; such suppression is termed primary suppression. For each primarily suppressed cell, other cells are selected and also subject to suppression (termed complementary suppression) in order to avoid disclosure of the primary cell count by algebraic operations of addition and subtraction on other unsuppressed cells and cell aggregates or marginal counts. The partner cells for complementary suppression are selected so that the *suppressed cell aggregate* obtained by aggregating the primary and its partner cells is disclosure-safe where some partner cells could themselves be subject to primary suppression. It is easily seen that the true count for safe suppressed cell aggregates can be deduced from other unsuppressed cells and cell aggregates. Thus, the cs-tabulation method has the desirable property of preserving the data integrity of all disclosure-safe cells and cell aggregates in that their true counts are either available directly from the unsuppressed information or can be deduced. It is also desirable in practice to choose aggregation partners for complementary suppression such that the resulting suppressed cell aggregate is substantively or analytically meaningful (i.e., construct-driven) in that there is a meaningful relationship to the extent possible between categories used for suppression partners. All suppressed cells in cs-tabulation are labeled ‘D’. Table 1(b) illustrates a choice of cells for complementary suppression. Using pre-specified preference rules for cell aggregation order such as the use of cells from the Other/Unknown category under race/ethnicity for complementary suppression and if that were not adequate, further or alternative aggregation over other race categories, and if that were not desirable, then the use of cells from categories under post-graduation plans, we designate the corresponding three cells in the Other/Unknown column as ‘D’ and a cell in the mixed race column in the government category as ‘D’. Thus there are a total of 10 cells out of 42 cells that are suppressed.

In any implementation of cs-tabulation, there are in general three concerns based on Duncan and Fienberg (1999) and Duncan et al. (2001).

1. Computationally Efficient and Meaningful Cell Aggregation for Complementary Suppression: For a given confidentiality threshold, it is computationally difficult in general for high dimensional tables to identify a set of cells for complementary suppression while allowing for exceptions such as below threshold counts for certain categories deemed disclosure-safe. Moreover, the resulting suppressed cell aggregates should be either analytically or substantively meaningful for user interpretability of a

tabular output with suppressed cells. For a practical implementation of cs-tabulation, there is need for a systematic efficient method for suppressed cell aggregation.

2. Valid Data Analysis and Uniform Interpretability: It is difficult to conduct a valid analysis in the presence of suppressed cells because different users may fill-in different values for suppressed cells from available information in unsuppressed counts of cells and cell aggregates using their own possibly ad hoc ways. This may lead to inconsistent conclusions in interpreting trends over different categories of a variable given other variables. For user convenience, enhanced data utility and uniform interpretability, there is a need by the data producer to fill-in suppressed cells by best estimates using only information contained in the unsuppressed cells and aggregates, and thus without increasing any disclosure risk posed by cs-tabulation. In principle, the same method could also be used by advanced users to fill-in suppressed cells from the current cs-tabular output. However, it would be preferable for the data producer to provide estimates for suppressed cells using a comprehensive and possibly complex method such that all released tables with common suppressed cells have common estimates.

3. Confidentiality Threshold Choice and Data Utility: If the choice of the confidentiality threshold is not data-specific, it is possible that it may not provide adequate disclosure safety for a given table. In other words, as part of a disclosure audit, it may be possible to "reverse-engineer" by estimating suppressed cells with very high precision based on information from unsuppressed cells and aggregates; this could happen because tabular data with unsuppressed cells and aggregates impose structural and algebraic constraints on suppressed cells subject to estimation. The threshold should be chosen such that estimated suppressed cells should neither be too close nor too far to true values in order to have a balance between data utility and confidentiality.

The cs-tabulation for SED is currently being implemented by using the τ-ARGUS software, originally developed in the 1990s at Statistics Netherlands (Hundepool et al., 2010), as an interface for entering tabular data into a mathematical optimizer to find complementary suppression cells. It is known that the problem of developing an efficient optimization algorithm for finding cells for complementary suppression in the context of high dimensional tables is mathematically difficult, innovations in τ-ARGUS have been made over the years. Although τ-ARGUS does not address the second and third concerns about cs-tabulation mentioned above, it does address the first concern but in a limited way. Since it uses an optimizing algorithm to find complementary suppressed cells, the resulting cell aggregates may not be substantively or analytically meaningful. It is limited to 3-dimensions and requires combining variables from a higher dimensional table in order to reduce it to 3-dimensions. Also it requires manual interventions to deal with suppression of zero cells and to deal with allowed exceptions such as cells with below threshold counts for certain categories deemed insensitive. In addition, it does not handle complementary suppression across multiple tables in a consistent manner especially if the number of tables is more than 10; therefore, manual review of cross-table suppressions is required to ensure that a cell that is suppressed in one table continues to be suppressed in other tables. A possible reason for this is that τ-ARGUS works with cells at the highest order of cross-classification and not with lower order cross-classifications; i.e., it uses a bottom-up approach, combined with the fact that there is not a unique way to finding complementary suppression partners.

The proposed method of ECS-Tab overcomes limitations of τ-ARGUS to a great extent and does address the three concerns in implementing cs-tabulation. ECS-Tab is based on the query-based PUF methodology of Singh, Borton, and Crego (2012); see also Singh et al. (2013). To address the first concern, ECS-Tab uses a quasi-hierarchical aggregation (qh-aggregation) of cells approach which is applicable to high dimensional tables and can be automated for ease in implementation. This approach uses a systematic

method to find disclosure-safe cells and cell aggregates in a top-down hierarchical manner starting with the grand total (0-dimension), and moving to each variable taken individually (1-dimension), and to each two-way combination of variables (2-dimension), and higher dimensional cross-classifications until the final highest dimensional table of interest is reached. At each stage, the prescribed confidentiality threshold is used to find primary and complementary suppression cells. Certain pre-specified hierarchical preference rules in cell aggregation order (between variables and between categories within a variable) are used to obtain substantively or analytically meaningful cell aggregates. Moreover, treatment of zero cells does not pose any new problems, and cells below the threshold that are allowed as exceptions can be easily accommodated. The qualifier 'quasi' is used to signify that categories of a variable selected as suppression partners given a category or a combination of categories of other variables may not be the same for other combinations. This is a departure from the usual hierarchical aggregation of cells because the goal of ECS-Tab is to preserve as many safe cells or aggregates as possible. Finally, when dealing with multiple tables corresponding to different user needs, cells suppressed in earlier tables that are common with the current table can be easily maintained as suppressed for consistency through a checklist.

Table 1(b) provides a simple illustration of qh-aggregation of cells under ECS-Tab. It shows how for each row category of the post-graduation plan, complementary suppression partners from the Other/Unknown column as well as from the mixed race column when needed can be assigned to primary suppression cells under the AI/AN and mixed race columns such that suppressed cell aggregates remain meaningful for user interpretation. Moreover, below threshold cells allowed as exceptions are not subject to suppression and therefore are not labeled 'D'. The cell aggregation here is hierarchical because 1-dimensional cells (i.e., row and column margins) are checked first for disclosure safety before 2-dimensional cells, and it is quasi because collapsing of cells from the two columns of AI/AN and Other/Unknown to obtain safe cell aggregates, for example, is not performed for all row categories. Although this is only a simple illustration of qh-aggregation performed manually, the approach is systematic by nature and can be easily computerized for higher dimensional tables as explained in Section 2 in the context of a more realistic 3-dimensional example.

To address the second concern of cs-tabulation, ECS-Tab provides optimal estimates of the counts of suppressed cells using a log linear model (Fienberg, 1980) under the assumption of multinomial sampling. Although it is natural for practical interpretability to use a regular hierarchical approach in selecting factor effects in log-linear modeling to obtain a parsimonious model with only significant effects, in ECS-Tab on the contrary, we fit a near saturated model by retaining as many estimable (or allowed) factor effects as possible. This is motivated from the key observation that selection of cells and cell aggregates that satisfy a minimum threshold is equivalent to collapsing of factor effects so that there is sufficient data to estimate them. Thus, the number of factor effects equals the number of safe unsuppressed cells and aggregates, and using only unsuppressed information, suppressed cells are estimated optimally such that all unsuppressed cells and aggregates are preserved at their true values. The above nontraditional approach to selecting factor effects is termed quasi-hierarchical, and is used to propose the qh-aggregation for ECS-Tab which performs cell aggregation in a top-down manner starting with 0-dim (analogous to inclusion of the intercept), 1-dim (analogous to main effects), 2-dim (corresponding to two factor effects) and so on. Regardless of the choice of suppressed cells being based on qh-aggregation or not, these cells can be estimated while preserving unsuppressed cells and cell aggregates. Also, any further modeling of the data to obtain a parsimonious model where the set of minimal sufficient statistics is obtained from the preserved cells and cell aggregates would be equivalent to modeling with the

original untreated table. Thus the resulting analysis remains valid regardless of the presence and estimation of suppressed cells. For a simple illustration of data utility due to the provision of estimates with ECS-Tab, Table 1(c) shows the best possible estimates under a nearly saturated model with 39 parameters out of a maximum of 42 (6 rows times 7 columns); the model has only 3 degrees of freedom which can be explained by the observation that although there are 10 suppressed cells, it is sufficient to know only 3 suppressed cells to deduce values of all other suppressed cells. From the estimates, the user can still elucidate the general trend in behavior over post-graduate plans for each of AI/AN, Mixed and Other/Unknown race/ethnicity categories. Section 3 contains a detailed description of estimation of suppressed cells.

To address the third concern of cs-tabulation, estimation of suppressed cells provides a built-in check of whether the corresponding confidentiality threshold provides an adequate protection from being able to fill-in suppressed cells with values too close to the true values; i.e., a disclosure audit. To this end, one can compute absolute error (AE) of estimated 'D' cells for primarily suppressed cells because they can be very small or even 0 and absolute relative error (ARE) for complementarily suppressed cells because they are at or above the confidentiality threshold, and check if median AE for primarily suppressed cells is not too small (such as using the cut-off value of 0.50--this choice is data dependent and is driven by the risk tolerance of the data producer), and median ARE for secondarily suppressed cells is also not too small using a cut-off value such as 5% that is not too large either in the interest of data utility. Table 1(c) shows errors in estimation of 10 'D' cells where errors in partner 'D' cells sum to zero as expected. If median AE or ARE is not large enough, we will need to increase the number of 'D' cells or decrease the number of unsuppressed cells and aggregates in order to introduce more uncertainty in estimated cells. This can be achieved by increasing the confidentiality threshold for lower dimensional tables in a top-down manner in order to propagate more 'D' cells at higher dimensions; i.e., by making the threshold dimension-specific. At each stage of increase in dimension, we can check using above error metrics whether there is sufficient uncertainty in estimated 'D', and if not, the threshold can be revised. Section 4 describes the disclosure audit process in detail.

The problem of maintaining internal consistency of tables from repeated user queries with respect to estimated cells is discussed in Section 5. It also includes summary and remarks and a brief comparison with alternative methods such as that of providing exact bounds for suppressed cells (Dobra and Fienberg, 2000) and controlled tabular adjustment of Dandekar and Cox (Cox et al., 2004).

2. Quasi-Hierarchical Cell Aggregation

We present a somewhat realistic 3-dimensional example to illustrate in a simple way that the steps underlying qh-aggregation are systematic, logical, and generalizable to obtain substantively or analytically meaningful cell aggregates for high dimensional tables. Thus qh-aggregation can add more utility to the table with suppressed cells than the current disclosure limitation methodology in SED even if estimates of suppressed cells are not provided under the proposed method of ECS-Tab. Suppose we have a 3-dimensional (2x3x4) table with variables: Sex, Citizenship, and Race/Ethnicity whose categories are shown in Table 2.1. For simplicity we are considering only two categories of sex, and four for race where White and Asian categories are collapsed together; this may be reasonable as they generally have similar distributions over other variable categories.

We will use the confidentiality threshold of 5 for cell suppression for disclosure-safety. Cells at risk along with suppression partners are suppressed or deleted. As

mentioned earlier, suppression of partner cells is equivalent to releasing values of their aggregate because the total count of partner cells in any suppression is at or above the threshold, and hence releasable.

For defining qh-aggregation, we need to specify preference order for choosing categories to be preserved in cell aggregation within a variable, and for choosing among variables in order of importance in that more important variables are considered for cell aggregation later than less important ones. Such preference rules are needed to have a direct control to ensure that resulting cell aggregates are analytically or substantively meaningful. Suppose the order of category aggregation preference between variables is taken as sex, citizenship, and race which is based on desired analysis requirements. That is, between citizenship and race, for example, category aggregation for race is first considered for each category of citizenship before considering category aggregation for citizenship in order to meet the confidentiality threshold via primary and complementary suppression. Thus, the variable deemed more important drives the order of category aggregation between variables and the final set of cell aggregates depends on the order specified.

We next specify preference rules for category aggregation within each variable. Since within sex, there are only two categories, there is only one choice of cell aggregation. For the citizenship variable, we will use Cit_3 (Other/Unknown) as a suppression partner if any one of Cit_1 (US Citizen) and Cit_2 (non-US Citizen) is at risk of disclosure. If the aggregate of both Cit_1 and Cit_2 is at risk, then all three categories are aggregated. Similarly, rules based on analysis goals for other possible scenarios of cell or category aggregation can be defined. In general, choice of aggregation partners should be based on a measure of similarity between categories if possible. If not, then it could be based on other considerations such as the similarity between distribution of counts over the levels of other variables or simply analysis goals. As a further example, for the race/ethnicity variable, we will use Rac_4 (Other) as a suppression partner if any one of the categories Rac_1, Rac_2, and Rac_3 is at risk. Such preference rules can be specified for other scenarios. For instance, Between Rac_1, Rac_2, and Rac_3, the order of preference for category preservation is Rac_1, followed by Rac_3, and then Rac_2. If the aggregate of Rac_2 and Rac_3 is at risk, for example, then use Rac_4 as their suppression partner. In the absence of any other alternative, the default option is to aggregate all the four citizenship categories.

To explain more clearly how the above basic preference rules within and between variables can be applied for qh-aggregation, consider a hypothetical 3-dimensional Table 2.2. There are several cells in Table 2.2 that are at risk. For disclosure-safety, the qh-aggregation is applied at different stages in a top-down hierarchical manner. We start in order with tables of 0-dim, 1-dim, 2-dim and so on. For a given dimension in the hierarchy, we consider cell aggregations necessary to satisfy the threshold under the hierarchy principle that all descendants of a suppressed cell (at a lower dimension) continue to be suppressed. The various stages of qh-aggregation for the above example are described below.

Stage 0: Start with the 0-dimensional marginal table which is simply the grand total. Clearly, this table is disclosure-safe using the confidentiality threshold of 5.

Stage 1: Check all 1-dimensional tables. Clearly, they are also disclosure-safe.

Stage 2: Now consider 2-dimensional tables where rows represent the variable with higher order of importance in cell aggregation than the column variable. The cells of the sex by citizenship table are all safe. However, such is not the case for the other 2-dimensional tables 2.3 and 2.4. Different notations are used for various types of suppressed cells as listed below. Such distinctions are useful to identify suppression partners under a pre-specified set of preference rules so that suppressed cell aggregates (i.e., total counts of suppression partner cells) become releasable. Denote by *D*: primary

suppression, *D'*: complementary suppression, *D''*: complementary-complementary suppression; i.e., complementary of complementary suppression, and *D'''*: descendant suppression under the hierarchy principle that if a cell in a lower dimensional table is suppressed, all its descendants in higher dimensional tables must also be suppressed.

In protecting an individual cell at risk, it can be seen that there are at least three other cells that are involved as suppression partners from which several safe cell aggregates can be released. In a 2-dimensional table, a suppressed cell is part of two safe cell aggregates—one in the horizontal direction across columns and the other in the vertical direction across rows. It may be noted that in some situations, other primary suppression (D) cells could serve as complementary cells (D') or complementary-complementary cells (D''). In finding complementary partners for any cell, we first look in the horizontal direction if the preferred variable as per the preference rule for cell aggregation is used as the row variable as was the case in Table 2.3, and then look in the vertical direction to select complementary suppression partners for both D and D' cells. If we change the order, the final choice of suppression partners may not be the same because choice of D' cells in one direction drives the choice of D' cells in the other direction. In practice, typically this is of consequence when we work with 3- or higher dimensional tables for finding suppression partners.

We still have to find the suppression partners for the 2-dimensional Table 2.4 of citizenship by race where the row variable is chosen as citizenship—the preferred variable over race. The choice of D' cells for the two D cells is based on preference rules. The two safe column cell aggregates (c2+c4) and the two safe row cell aggregates (r1+r2) can be released as explained for Table 2.3. Notice that under qh-aggregation, aggregation of categories for one variable is not performed uniformly for all categories of the other variable. For example, in the table 2.4, columns 2 and 4 are aggregated for rows 1 and 2 but not for row 3. This feature is desirable for minimizing suppression and, in fact, is the reason for using the prefix 'quasi' to distinguish it from the usual hierarchical aggregation where if at any stage, a category of a variable is aggregated with another category, it remains collapsed for all higher stages; i.e., the two categories can no longer be separated.

Stage 3: For the 3-dimensional table of sex by citizenship by race (Table 2.5), we cast the aggregated version of the 2-dimensional table (sex by citizenship) obtained in Stage 2 as rows of a new 2-diemsional table while the third variable race as columns in view of the convention that the row variables (sex and citizenship) have the preference in order of importance for cell aggregation over the race variable.

In the final Table 2.6 after suppression, although there are 16 suppressed elementary cells in the sex by citizenship by race table and 8 suppressed margins (4 in the citizenship by race table and 4 in the sex by race table) with a total of 24 suppressed cells, there is only 5 effective number of suppressed cells in that if we know 5 suppressed cells suitably chosen, then we can obtain values of all other suppressed cells using values of unsuppressed cells and cell margins. To see this, consider the first marginal 2-dim table of citizenship by race which has 4 suppressed cells while the effective number of suppressed cells is only 1 (the dark shaded cell). Similarly, for the marginal table of sex by race with 4 suppressed cells, there is effectively only 1 suppressed cell (dark shaded). Now among the 16 suppressed cells in the sex by citizenship by race table, first consider the subtable for Sex_1 with 8 suppressed cells. If we know values for 3 cells (shown in the dark shade), along with all cells in the marginal tables, we can deduce the value of cell (Sex_1,Cit_1,Rac_2) from values of cells (Sex_1,Cit_2,Rac_2), (Sex_1,Cit_3, Rac_2) and the margin cell (Sex_1,Rac_2). Next, the value of cell (Sex_1,Cit_1,Rac_2) discloses the value of cell (Sex_1,Cit_1,Rac_4) because of known margin (Sex_1,Rac_1) and other cells. Similarly, the value of cell (Sex_1,Cit_3, Rac_4) can be deduced which, in turn, discloses the value of cell

(Sex_1,Cit_2,Rac_4), and finally that of cell (Sex_1,Cit_2,Rac_3). In an analogous manner, values of 8 suppressed cells in the second subtable for Sex_2 can be easily deduced from the known margin of citizenship by race and the disclosed and known cell values in the subtable Sex_1. Thus we have a total of 5 effective number of suppressed cells; 1 from each of the two 2-dim margins, and 3 from the full 3-dim table. This completes the description of the process to obtain the final table under qh-aggregation.

3. Estimation of Suppressed Cell Counts and Its Impact on Analysis

The utility of tables with suppressed cells can be enhanced by replacing suppressed cell counts with estimates. Replacing suppressed cells with estimates gives the user an indication of the magnitude of the suppressed counts and the underlying trend over different categories for a given subpopulation or domain of interest. Additionally, this method can keep the true value of each suppressed cell adequately uncertain, depending on the model used for estimation. This is important because the estimates are based solely on the unsuppressed information within a table, or family of tables, and so could also be produced by an advanced data user. However, it is preferable for the data producer to provide estimates. This allows the producer to introduce adequate uncertainty in the estimates. It is also more convenient and useful for data users as it allows the data producer to employ complex models for the best estimation and to maintain consistency across tables. The estimation technique discussed here can be implemented regardless of what method was used to determine complementarily suppressed cells (e.g. qh-aggregation or a non-qh-aggregation under a mathematical optimizer as in τ-ARGUS).

The method ECS-Tab proposed here for estimating suppressed cells uses hierarchical log-linear modeling (Fienberg, 1980) to estimate model parameters corresponding to the constraints on estimated counts for all cells (not just suppressed cells) given by cells and cell aggregates being preserved. This modeling is not standard because parameters do not correspond to usual main effects, and lower and higher order interactions. However, with the proposed qh-aggregation approach for cell aggregation for finding suppressed cells, the interpretation of parameters is somewhat analogous to the usual factor effects under a hierarchical modeling approach.

In a hierarchical modeling approach, lower order factor effects are included before higher order ones. Additionally, if certain factor levels are collapsed (like cell aggregation for reasons of insufficient data) for any given factor effect in the hierarchical order, then all factor effects in that order respect the same level of collapsing. In other words, if at any stage, a category of a variable is aggregated with another category, it remains collapsed for all higher stages; i.e., the two categories can no longer be separated. This is, however, not done under a qh-aggregation approach, and is desirable for minimizing suppression; hence the reason for using the prefix ‘quasi’ to distinguish it from the usual hierarchical aggregation. Thus, under qh-aggregation, model parameters do not have a natural interpretation in terms of factor effects. This is fine for our purposes because the goal of ECS-Tab is not to find the most parsimonious and practically meaningful model but to find the least parsimonious (near saturated) model so that there are as many parameters as the number of linearly independent constraints or safe cells and cell aggregates. The following section provides a technical description of estimating suppressed cells under ECS-Tab which may be omitted at first reading.

Regardless of how the suppressed cells were obtained (quasi-hierarchically under a top-down approach or non-quasi-hierarchically under a bottom-up approach) for a given cross-classified table, one can find a set of linearly independent constraints in terms of cells and cell aggregates that must be satisfied by estimated counts under the model. To do this, one can form a matrix with rows having elements of 1s and 0s with the number of elements being equal to the total number of cells in the final cross-classified table of

interest; i.e., elementary cells, say *M*, and where rows correspond to all constraints of cells and cell aggregates. Each row has a 1 in the place that indicates which cell is included in the constraint, and 0 elsewhere. It is possible that all cells are labeled 'D' and constraints are only in terms of marginals or cell aggregates. Rows of this matrix could be linearly dependent because some constraints could be derived from others by algebraic manipulations. However, it is sufficient to work with only linearly independent constraints. So we reduce the row dimension of the matrix to achieve linear independence. Suppose there are *p* independent rows and the number of columns is *M*—the total number of cells in the cross-classified table of interest. The estimating equations for *p* model parameters ($\beta's$) can be written as follows.

$$\begin{pmatrix} x_{11}\, x_{21} \ldots x_{k1} \ldots x_{M1} \\ x_{12}\ x_{22} \ldots x_{k2} \ldots x_{M2} \\ \ldots \\ \ldots \\ x_{1i}\ x_{2i} \ldots x_{ki} \ldots x_{Mi} \\ \ldots \\ \ldots \\ x_{1p}\ x_{2p} \ldots x_{kp} \ldots x_{Mp} \end{pmatrix} \times \begin{pmatrix} \exp(x_{11}\beta_1 + x_{12}\beta_2 + \cdots x_{1i}\beta_i + \cdots x_{1p}\beta_p) \\ \exp(x_{21}\beta_1 + x_{22}\beta_2 + \cdots x_{2i}\beta_i + \cdots x_{2p}\beta_p) \\ \ldots \\ \ldots \\ \exp(x_{k1}\beta_1 + x_{k2}\beta_2 + \cdots x_{ki}\beta_i + \cdots x_{kp}\beta_p) \\ \ldots \\ \ldots \\ \exp(x_{M1}\beta_1 + x_{M2}\beta_2 + \cdots x_{Mi}\beta_i + \cdots x_{Mp}\beta_p) \end{pmatrix} = \begin{pmatrix} t_1 \\ t_2 \\ . \\ . \\ t_i \\ . \\ . \\ t_p \end{pmatrix}$$

where log of the expected counts are assumed to be linear in $\beta's$, t_i denotes the *i*th nonnegative constraint (value of safe cell or cell aggregate), and x_{ki} is the *k*th column element of the *i*th row of the constraint matrix taking values of 1 or 0; *i*=1,…, p; *k*= 1,…,M. There are *p* equations and *p* unknowns ($\beta's$) and the above system of nonlinear equations in principle can be solved by well known methods such as Newton-Raphson. However, in real applications, *p* can be quite large (in tens of thousands) and so an alternative method of nonlinear Gauss-Seidel (Jiang, 2000) can be used. It may be noted that although all cell counts are estimated (i.e., both suppressed and non-suppressed cells) and hence all cell aggregates, cells and cell aggregates that are safe are preserved at their true values. In other words, estimated counts match safe cell and cell aggregate values (to be preserved) by construction via constraints in the estimating equations. The above estimating equations coincide with maximum likelihood equations for the model parameters under the assumption of multinomial sampling, and thus the estimates are optimal.

The table with estimated suppressed cells using ECS-Tab allows for valid data analysis in general. In particular, for descriptive analysis including trend comparisons, it provides easily interpretable point estimates. However, for variance estimation of point estimates, more work is needed to account for estimated suppressed cells counts. Usual hierarchical modeling of the tabular data can proceed as in the case of the original table as long as the sufficient statistics used in modeling are based on safe cells and cell aggregates obtained under qh-aggregation. The reason for this is that under hierarchical modeling as mentioned earlier, in defining two and higher order factor effects, collapsing of categories for a given variable is done uniformly over all the combinations of categories of other variables unlike qh-aggregation. So, sufficient statistics for estimating corresponding factor effects can be obtained by further collapsing the unsuppressed cells and cell aggregates obtained under the quasi hierarchical aggregation.

4. Disclosure Audit

An important feature of the proposed ECS-Tab method is that it provides a built-in means for disclosure audit, which is not provided under the current method of cs-

tabulation. ECS-Tab automatically conducts a disclosure audit because we can compute AE or ARE of estimated counts for suppressed cells with respect to their true values to check if there is sufficient uncertainty about the true value. Without sufficient uncertainty, a suppressed cell might still be disclosure-prone due to precise estimation. In other words, this gives a metric of disclosure risk. In practice, it might be better to define disclosure risk separately for two types of suppressed cells—primary and complementary suppression because the ARE threshold for complementary suppressed cells can be set small due to true cell counts being above the confidentiality threshold. For each type of suppressed cells, quantiles of the ARE distribution can be used as measures of disclosure risk. Disclosure risk in terms of the amount of uncertainty required for adequate protection must be defined in advance by the data producer. For suppressed cells with zero or very small counts, it is more meaningful to use AE instead of ARE..

The goal in any disclosure treatment process is to balance data utility against disclosure risk. So while we do not wish ARE to be too small for a given suppressed cell for disclosure-safety, we do not want it to be too large either for key analysis domains; i.e., combinations of selected cells. Suppose for a given cross-classified table, and a pre-specified choice of confidentiality threshold (such as 5), estimates for certain suppressed cells using the ECS-tab method turn out to be too close to the true values. How can we introduce more error? In general, the more the parameters in the ECS-Tab model, i.e., the more the number of cells and cell aggregates being preserved, the more accurate the estimates will be because they will have less room for fluctuation under the constraints to be satisfied. For example, for 3-dimensional tables, we can model directly each table to obtain estimates for suppressed cells. Alternatively, we could first consider a corresponding higher dimensional table such that the 3-dimensional table of interest is obtained as a margin (or sub-table). We would then identify suppressed cells for the higher dimensional table using the prescribed threshold, and fit a model to estimate these cells which, in turn, would provide estimates for suppressed cells in the 3-dimensional sub-table. For a given threshold, we would expect more accurate estimates if the model is fit for a higher dimensional table from which the desired lower dimensional table with estimated counts is obtained.

There are advantages in working with higher dimensional tables. In particular, models are fit once for each such table and then all sub-tables of interest can be obtained without further modeling. If we want to introduce more error in our estimates, there are several strategies we could employ. Under the qh-aggregation, one strategy might be to increase the confidentiality threshold for lower dimensional tables in order to allow for more suppression and decrease the threshold as the dimension increases to reduce over-suppression at higher dimensions. It follows from the above discussion on the impact of increasing threshold at lower dimensions that in practice, a useful strategy might be to let the choice of the confidentiality threshold for each dimension in the qh- aggregation be data-driven. Specifically, assume without loss of generality that the grand total is safe. Then, look at 1-dimensional margins, and using an initial large threshold (such as 10) for unsafe cells, find suppressed cells by choosing a threshold for each 1-dimensional table separately such that AREs for primarily suppressed cells after modeling are not too small. Similarly, we would choose thresholds for 2-dimensional tables except that estimates of suppressed cells obtained at the previous stage are preserved; i.e., modeling is used to find a suitable threshold for new cell suppression such that AREs are not too small. Next we go to 3-dimensional tables and so on until we reach the final table of interest. It might be better to start modeling at dimension three if the table of interest is at least 3-dimensional, to get reasonable estimates of suppressed cells so that AREs are not too large either. These and other variants of developing a suitable strategy for choosing confidentiality thresholds can be part of any future development of the ECS-Tab

methodology so that a balance between data utility against disclosure risk can be achieved.

5. Concluding Remarks

We remark that in any practical implementation of ECS-Tab, it is important in the interest of internal consistency to ensure that estimates of cells released in earlier queries are preserved in later queries for tabular output of cells that are common. This can be done by constructing a checklist of variables with allowable categories for one-dimensional, two-dimensional, and higher order marginal distributions. We need two Checklists I and II: the first list for allowable cell aggregates which could be cells by themselves, and the second list for non-allowable or suppressed cells. The internal consistency is accomplished by enlarging the vector of safe cell or cell aggregates in modeling to include estimates obtained for the first time for such aggregates. Using the checklist of allowable cell aggregates for pre-screening, the query-based PUF on which ECS-Tab is based can thwart differencing attacks—a very difficult problem for query-based systems as discussed in Gomatnam et al. (2005).

Below we list the main modules required by ECS-Tab in any application.

1. *Data Set-up* (for selecting variables and representing cells and cell aggregates stacked in a column; this is useful later on for modeling because a traditional tabular form with cells in high dimensions is less tractable for finding safe cell aggregates)
2. *Quasi-hierarchical Cell Aggregation* (for constructing a column of disclosure-safe cell aggregates)
3. *Checklists I and II* (corresponding to each entry in the column of cell aggregates, List I consists of safe or publishable cells and cell aggregates while List II consists of suppressed cells including both primary and complementary)
4. *Cell Aggregate Column Enlargement* (for including past estimates for common suppressed cells in order to ensure consistency between estimates from past and current tables)
5. *Model Specification and Fitting* (for estimating suppressed cells using log linear models such that all cells and cell aggregates in the column of cell aggregates from Module 4 are exactly satisfied by the table of disclosure-safe cell counts and estimates of suppressed cell counts)
6. *Disclosure Audit* (for ensuring estimates of suppressed cell counts are not too close to the true values; else consider revisions of the disclosure threshold corresponding to each stage in the hierarchy of dimensions))

We conclude with a brief discussion of other alternatives to cs-tabulation. The method of controlled tabular adjustment (CTA) of Dandekar-Cox (Cox et al., 2004) can be viewed as a form of output disclosure treatment. In CTA, all sensitive cells defined by the threshold criterion are replaced by tolerance limits (upper or lower in the case of magnitude data, but in the case of count data, these are simply the prescribed minimum threshold count) and then the whole table is perturbed using linear programming to preserve various marginal counts and other constraints. By contrast, in ECS-Tab, only small and complementary cells are perturbed by way of estimation based on a data-driven model and not on an external mechanism. Another method due to Dobra and Fienberg (2000; see also, Fienberg, 2001) provides exact sharp bounds for 'D' cells based on released cells and cell aggregates, and is an alternative to estimating 'D' cells. The bounds are not based on probability considerations and their tightness depends on the threshold criterion. This is an interesting alternative as it recognizes explicitly the uncertainty about the values of 'D' cells and could be used as a guide in choosing

appropriate threshold for disclosure-safety. However, in practice, the bounds are typically quite loose, and users might prefer instead point estimates (with possibly associated standard errors) as provided by ECS-Tab for ease in interpretation

Disclaimer and Acknowledgments

The opinions expressed in this paper represent the views of the authors and not the official position of the National Science Foundation. The authors thank Mark Fiegener and Darius Singpurwalla of NSF for useful comments, Mary Ann Latter, Dan Kasprzyk and Bronwyn Nichols of NORC for their support and encouragement and especially Stephen Schacht of NORC for useful discussions and feedback.

References

Cox, L.H. (1980). Suppression methodology and statistical disclosure control. *Journal of the American Statistical Association*, 75, pp. 377-85.

Cox, L. H., Kelly, J. P. and Patil, R. (2004). Balancing quality and confidentiality for multivariate tabular data. *Lecture Notes in Computer Science*, 3050, 87–98.

Dobra, A. and Fienberg, S.E. (2000). Bounds for cell entries in contingency tables given marginal totals and decomposable graphs. *Proceedings of the National Academy of Sciences*, 97, 11885-11892.

Duncan, G.T., Fienberg, S.E., Krishnan, R., Padman, R., Roehrig, S.F. (2001). Disclosure limitation methods and Information loss for tabular data. In *Confidentiality, Disclosure, and Data Access, Eds. Doyle, P, Lane, J.I., Theeuwes, J.J.M. and Zayatz, L.V.*, North-Holland, Elsevier, New York.

Duncan, G.T., and Fienberg, S.E. (1999). Obtaining information while preserving privacy: A Markov perturbation method for tabular data. In Statistical Data Protection (SDP'98)Proceedings, Luxembourg, Eurostat, pp. 351-62. http://www.heinz.cmu.edu/research/21full.pdf

Fienberg, S.E. (1980). The Analysis of Cross-classified Categorical Data. MIT Press, Cambridge, MA.

Fienberg, S.E. (2001). Statistical perspectives on confidentiality and data access in public health. Statistics in Medicine, 20, pp. 1347-56.

Fischetti, M., and Salazar, J.J. (2000). Models and algorithms for optimizing cell suppression in tabular data with linear constraints. *Journal of the American Statistical Association, 95, pp. 916-28.*

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Naylor, J., Nordholt, E.S., Seri, G, and De Wolf, P-P. (2010), Handbook on Statistical Disclosure Control, Version 1.2, ESSNet publication, http://neon.vb.cbs.nl/casc/handbook.htm

Jiang, J. (2000). A nonlinear Gauss-Seidel algorithm for inference about GLMM. *Computational Statistics*, 15, 229-241.

Singh, A.C., Borton, J.M., Davern, M. E., and Lin, Y. (2013). Query-based PUFs for Disclosure-Safe Remote Analysis from Medicare Claims Micro Data, *ASA Proc., Surv. Res. Meth. Sec.*,

Singh, A.C, Borton, J.M., and Crego, A.M. (2012). A generalized domain size threshold for analysis restrictions with remote analysis servers. Proceedings of the Federal Committee on Statistical Methodology, US Census Bureau. http://www.fcsm.gov/events/papers2012.html

Table 1(a): Post-graduation plans of doctorate recipients in Life Sciences in the US using a hypothetical data based on SED

Post-Graduation Plan	AI/ AN	Asian	Black	Hisp	White	Mixed (Two or More)	Other/ Unk		Row Total
Postgrad Study or training	7	329	117	197	2348	78	38		3114
Academe	2	42	40	34	638	10	12		778
Industry/ Business	2	47	14	6	251	4	2		326
Gov	1	16	24	10	175	6	2		234
Non-profit	0	16	10	5	100	1	4		136
Other/Unk	0	1	4	1	38	0	0		44
Column Total	12	451	209	253	3550	99	58		4632

Table 1(b): Post-graduation plans of doctorate recipients in Life Sciences in the US using a hypothetical data based on SED ('D' Cells using threshold of 5)

Post-Graduation Plan	AI/ AN	Asian	Black	Hisp	White	Mixed (Two or More)	Other/ Unk		Row Total
Postgrad Study or training	7	329	117	197	2348	78	38		3114
Academe	D	42	40	34	638	10	D		778
Industry/ Business	D	47	14	6	251	D	2		326
Gov	D	16	24	10	175	D	D		234
Non-profit	D	16	10	5	100	D	D		136
Other/Unk	0	1	4	1	38	0	0		44
Column Total	12	451	209	253	3550	99	58		4632

Table 1(c): Post-graduation plans of doctorate recipients in Life Sciences in the US using a hypothetical data based on SED (Estimated 'D' cells, Errors shown underneath)

Post-Graduation Plan	AI/ AN	Asian	Black	Hisp	White	Mixed (Two or More)	Other/ Unk		Row Total
Postgrad Study or training	7	329	117	197	2348	78	38		3114
Academe	2.48 +.48	42	40	34	638	10	11.52 -.48		778
Industry/ Business	1.12 -.88	47	14	6	251	4.88 +.88	2		326
Gov	.90 -.10	16	24	10	175	3.94 -2.06	4.16 +2.16		234

Non-profit	.50 +.50	16	10	5	100	2.18 +1.18	2.32 -1.68		136
Other/Unk	0	1	4	1	38	0	0		44
Column Total	12	451	209	253	3550	99	58		4632

Table 2.1: Variable Category Definitions for the Example

Variable	Categories
Sex	Sex_1: Female, Sex_2:Male/Unknown
Citizenship	Cit_1: US Citizen/Permanent Resident, Cit_2: Non-US Citizen/Permanent Resident, Cit_3: Other (temporary visa holders)/Unknown
Race/Ethnicity	Rac_1:White/Asian, Rac_2:Hispanic, Rac_3:Black (non-Hispanic), Rac_4:Other (American Indian/Alaska Native)/Two or More Races/Unknown

Table 2.2: A Hypothetical 3-Dimensional Table

3-Dim	Total						Sex_1					Sex_2				
	All	Rac_1	Rac_2	Rac_3	Rac_4		All	Rac_1	Rac_2	Rac_3	Rac_4	All	Rac_1	Rac_2	Rac_3	Rac_4
All	108	33	12	29	34		49	16	4	13	16	59	17	8	16	18
Cit_1	33	11	1	10	11		15	5	0	5	5	18	6	1	5	6
Cit_2	38	10	4	11	13		17	5	1	5	6	18	5	3	6	7
Cit_3	37	12	7	8	10		17	6	3	3	5	20	6	4	5	5

Table 2.3: Margin of Sex by Race/Ethnicity

2-Dim	Rac_1	Rac_2	Rac_3	Rac_4	All
Sex_1	16	**4 (D)** c2+c4 r1+r2	13	**16 (D')** c4+c2 r1+r2	49
Sex_2	17	**8 (D')** c2+c4 r2+r1	16	**18 (D'')** c4+c2 r2+r1	59
All	33	12	29	34	108

Footnote to Table 2.3a: D denotes primary suppression, D' denotes complementary suppression, and D'' denotes complementary-complementary suppression. The entry of c2+c4 in the cell (Sex_1, Rac_2), for example, indicates that cells in columns 2 and 4 of row 1 are suppression partners for a safe cell aggregate. Similarly, r1+r2 signifies that cells in rows 1 and 2 of column 2 are suppression partners.

Table 2.4: Margin of Citizenship by Race/Ethnicity

2-Dim	Rac_1	Rac_2	Rac_3	Rac_4	All
Cit_1	11	**1 (D)** c2+c4 r1+r2	10	**11 (D')** c4+c2 r1+r2	33
Cit_2	10	**4 (D)** c2+c4 r2+r1	11	**13 (D')** c4+c2 r2+r1	38
Cit_3	12	7	8	10	37
All	33	12	29	34	108

Table 2.5: Full Table of Sex by Citizenship by Race/Ethnicity

3-Dim	Rac_1	Rac_2	Rac_3	Rac_4	All
Sex_1,Cit_1	5	0 (D''') c2+c4 r1+r2+r3+…	5	5 (D''') c4+c2 r1+r2+r3+…	15
Sex_1,Cit_2	5	1 (D''') c2+c3+c4 r1+r2+r3+…	**5 (D')** c3+c2+c4 r2+r3	6 (D''') c4+c2+c3 r1+r2+r3+…	17
Sex_1,Cit_3	6	3 (D''') c2+c3+c4 r1+r2+r3+…	**3 (D)** c3+c2+c4 r3+r2	5 (D''') c4+c2+c3 r1+r2+r3+…	17
Sex_2,Cit_1	6	1 (D''') c2+c4 r4+r5+r6+…	5	6 (D''') c4+c2 r4+r5+r6+…	18
Sex_2,Cit_2	5	3 (D''') c2+c3+c4 r4+r5+r6+…	**6 (D')** c3+c2+c4 r5+r6	7 (D''') c4+c2+c3 r4+r5+r6+…	21
Sex_2,Cit_3	6	4 (D''') c2+c3+c4 r4+r5+r6+…	**5 (D')** c3+c2+c4 r6+r5	5 (D''') c4+c2+c3 r4+r5+r6+…	20
All	33	12	**29**	34	108

Footnote: D''' denotes descendant suppression cells under the hierarchy principle because of suppression of the parent cell. The corresponding safe cell aggregate in the value released (e.g., r1+r2+r3+…) also includes suppression partners from the lower dimensional table of sex by race where the parent cell was suppressed.

Table 2.6: Full Table after Suppression under qh-Aggregation (Threshold of 5)

3-Dim	Total						Sex_1						Sex_2				
	All	Rac_1	Rac_2	Rac_3	Rac_4		All	Rac_1	Rac_2	Rac_3	Rac_4		All	Rac_1	Rac_2	Rac_3	Rac_4
All	108	33	12	29	34		49	16	**4 D**	13	**16 D'**		59	17	**8 D'**	16	**18 D''**
Cit_1	33	11	**1 D**	10	**11 D'**		15	5	0 D'''	5	5 D'''		18	6	1 D'''	5	6 D'''
Cit_2	38	10	**4 D**	11	**13 D'**		17	5	1 D'''	**5 D'**	6 D'''		18	5	3 D'''	**6 D''**	7 D'''
Cit_3	37	12	7	8	10		17	6	3 D'''	**3 D**	5 D'''		20	6	4 D'''	**5 D'**	5 D'''

Footnote: Under the Top-Down approach, start with 0-dim, 1-dim, 2-dim… tables in order to find D cells for primary suppression, D' cells for complementary suppression, D'' cells for complementary-complementary suppression, and D''' cells for descendant suppression.