MULTIDIMENSIONAL TABULAR DATA SUPPRESSION IN THE CANADIAN CANCER REGISTRY

Eric Hortop[1]

ABSTRACT

The Canadian Cancer Registry is a national database of cancer cases compiled from provincial and territorial cancer registry data. Among the tabulations publicly available from the registry are tables of incidence, which are cross-tabulated by geography, tumour site, sex and age. These high-dimensional tables of rare events are available via Statistics Canada's CANSIM II system and require a disclosure control strategy. This paper presents a system that counters a hacker with some subject matter knowledge by iteratively suppressing cells, then verifying table security using a separately defined mathematical programming problem. It also provides graphical and tabular process information to analysts.

KEY WORDS: Confidentiality, Incidence, Rare events, Suppression

RÉSUMÉ

Le Registre canadien du cancer est une base de données nationale des cas de cancer recueillis auprès des registres provinciaux et territoriaux sur le cancer. Parmi les tableaux du registre diffusés au grand public, on trouve des tableaux d'incidences classés selon la géographie, l'organe atteint par la tumeur, le sexe et l'âge. Ces tableaux multidimensionnels des événements rares peuvent être consultés dans le système CANSIM II de Statistique Canada et nécessitent une stratégie de contrôle de la divulgation. Cette communication présente un système qui assure une protection contre un intrus ayant des connaissances du domaine spécialisé au moyen d'une suppression itérative des cellules, pour vérifier ensuite la sécurité des tableaux en utilisant un problème de programmation mathématique défini séparément. En outre, il fournit aux analystes un résumé graphique et tabulaire du processus.

MOTS CLÉS : Confidentialité; événements rares; fréquence; suppression

1 Statistics Canada, 16th floor, 100 Tunney's Pasture Driveway, Canada, K1A 0T6, eric.hortop@statcan.gc.ca

1. INTRODUCTION

1.1 The Cancer Incidence Suppression Problem

In 2008, Statistics Canada added cross-tabulations of cancer incidence cross-tabulated by more classification variables than before to its publicly-accessible CANSIM II database. The new tables contained crude rates by age, sex, geography and site of tumour (Statistics Canada, 2009a), and age-standardized rates by sex, geography and site of tumour (Statistics Canada, 2009b). The preceding tables had only two dimensions, and were protected against residual disclosure of individuals and small counts by cell-suppression software (Merrigan, 2003) not designed for higher-dimensional tables. The new tables were to be protected by cell suppression, so the algorithm would need to be updated to accommodate tables of higher dimension. The new system is dubbed "Tau-Ardent," in reference to the similarly-named system from Statistics Netherlands, but "ardent" for its speed at the cost of optimality and its single-pass methodology.

A sensitive cell is defined in the Canadian Cancer Registry (CCR) case as cells with counts ranging from one to five cases. Tabulating cancer incidence rates by site and sex, as was required for both tables, introduced additional complication in that some incidence rates of zero are in "impossible" cells for tumour sites which are sex-specific. In order to publish as many cases' cells as possible, empty-but-biologically-plausible cells were considered candidates for cell suppression, but cells which an attacker could assume to be zero from their sex and site (prostate cancer in women or cervical cancer in men, for example) and cells whose data is known to be missing (some jurisdictions lag behind others in submitting their cases to the CCR) provide no protection because an attacker can simply infer their value to be zero, and possibly progress to calculating the values of sensitive cells sharing a sex or tumour site with the "impossible" cell, so they were not candidates for suppression. Another issue with the subject matter, applicable only to crude rates, was that cancer occurrence is strongly linked to age, so suppression of an empty cell provides more uncertainty as to its actual count when its age co-ordinate is consistent with the occurrence of cancer. Suppressing an implausible empty cell provides little or no protection, so an automated way of favouring plausible cells is very desirable if empty cells are candidates for suppression.

The incidence tables for all years are regenerated every year from 1992 to the latest data available to allow for late submissions and revised records, and the larger, crude table contains over 45 000 cells per year and nearly 10 000 sensitive cells, so minimal manual intervention and extensive summary information on the suppression process were very important.

1.2 General Cell-Suppression Problem for Counts

The goal of cell suppression is to prevent the overly-accurate estimation of sensitive cells. Ways to prevent exact counts of small cells being released include random rounding, collapsing of categories, perturbation of underlying microdata and cell suppression (Willenborg and de Waal, 1996). In the cell-suppression case, all sensitive cells' values are suppressed. Unfortunately, simple suppression of sensitive cells in tables with marginal totals can be ineffective if an attacker can solve for the value of a missing cell using unsuppressed interior cells and marginal cells. For example, if only one cell is suppressed in a column, all the other values in the column can be subtracted from the column total, yielding the suppressed cell's value. In order to prevent the disclosure of sensitive cells, a cell suppression strategy must also be able to suppress non-sensitive cells to impede the calculation of sensitive cells' values: this is called "complementary suppression". The released data from the Canadian Cancer Registry does not distinguish between sensitive cells and cells suppressed by complementary suppression; any given cell is either visible or suppressed.

In a tabular cell-suppression context, the degree of protection given to a cell can be measured by a "feasibility interval" (Willenborg and de Waal, 1996, pp. 97–101): the range of possible values for a suppressed cell, given that all cell values are non-negative integers and using information from marginal and unsuppressed internal cells. In the CCR case, feasibility intervals are not published (although they can be calculated without using sensitive information). For CCR suppression, the feasibility interval of any suppressed cell had to contain at least two integers. In practice, feasibility intervals were typically significantly wider than the minimum.

2. CELL-SUPPRESSION APPROACHES

2.1 The Merrigan Suppression Program

The original CCR suppression program worked iteratively in four passes. First it marked all small-count cells for suppression in one pass. The two subsequent passes examined rows and columns, suppressing the first visible non-zero cell having the row (or column) minimum value in any row (or column) having only a single suppression. The final pass re-examined rows; any row with a lone suppression had the first visible non-zero cell in a column already having at least one suppressed cell. Cells flagged in the last pass could be large cells; manual verification was recommended for them. The new program was to be an update of the Merrigan approach, keeping its general behaviour but working on three- and four- dimensional tables as well as requiring less manual verification.

2.2 The Hypercube Method

The hypercube method of secondary cell suppression (Willenborg and de Waal, 1996, pp 119–122) is a method of cell suppression that works by forming a "best" hypercube of 2^D cells to suppress with each sensitive cell in turn, and suppressing any visible cells at the points of that hypercube (it is possible – and desirable – for the chosen hypercube to be composed entirely of suppressed cells). The desirability of suppressing a cell is determined by a weight representing how much information is lost by suppressing that cell – Willenborg and de Waal leave the calculation of that weight open to interpretation, but frequency count in that cell is a possibility. A suppressed cell should be assigned a large negative weight to encourage its inclusion in subsequent hypercubes. The hypercube method is a simple, effective method that Willenborg and de Waal take as a baseline, which illustrates the need to protect complementary suppressions as well as actual sensitive cells.

The hypercube method was not used for cancer incidence tabulations in part because the Merrigan program worked in a very different way, and the original intent was to modify the Merrigan program to support extra dimensions rather than build a new cell suppression system from the ground up. When tested on small datasets, the hypercube method as implemented by τ Argus 3.3 (Statistics Netherlands, 2008) suppresses more cells than the algorithm described in this paper.

2.3 CONFID2

Statistics Canada has a generalized confidentiality system for tabular data, CONFID2, described in Frovola, Fillion and Tambay, 2009. It was released after development of the CCR's confidentiality software began, and addresses a slightly different disclosure control problem. CONFID2 provides an optimized suppression pattern based on cell sensitivity — for each sensitive cell, it finds optimal complementary suppressions based on a user-chosen score function, and then it refines its suppression pattern by finding the best adequate subset of complementary suppressions based on a second user-chosen score function. Complementary suppressions must include cells of sufficient size to cover the sensitivity calculated for the sensitive cell. Sensitivity is in turn calculated based on dominance: dominance occurs when the top contributors to a cell make up too much of the cell total, for some definition of "too much". This is an accepted and useful definition for magnitude data, like income, production volumes and land area.

The first problem with using CONFID2 for cancer incidence tables is that it is designed for magnitude data, where respondents contribute different amounts to their respective cells.

The other problem is that CONFID2 will only suppress non-zero cells. In cancer data, cells with low underlying rates can be zero or have a small count, and in order to protect as many higher-incidence cells as possible, the subject-matter experts on the project wanted to allow the suppression of zero cells where they were not zero by definition.

3. THE NEW CANCER INCIDENCE IMPLEMENTATION: TAU-ARDENT

3.1 A Basic Iterative Method

As a starting point, we describe an iterative method for finding complementary suppression. We start with a table of dimension *D*, with marginal totals included and no hierarchy beyond marginal totals and interior cells. We suppress cells one at a time. We define a "problem" as a pair of a suppressed cell and a direction (a variable) where the only value of the specified variable for which a cell matching the suppressed cell on all other variables is the one matching the suppressed cell. One cell may entail as many as *D* problems. We define a "problem lane" to be the set of cells, identified by and summing to a marginal cell, associated with a problem where there is exactly one suppressed cell.

For all unsuppressed cells in the table, we calculate "problems solved" *s*: the number of problem lanes including the unsuppressed cell, and "lowest directions" *l*: number of directions in which the cell has the minimum count. We calculate a score $S = (D+1)s + l$ for each cell with $s > 0$, and $S=0$ for each cell with $s = 0$. We then calculate S^*, an adjusted score from *S*. We halve S for marginal cells to reflect a desire to keep them. We then assign an arbitrary size penalty based on cell size (using a monotonically nondecreasing step function), to fine-tune the suppression pattern and reflect a desire to keep large cells. The subtraction of the penalty is subject to the restriction that S^* must be strictly greater than zero if *S* is greater than zero — if S^* is reduced too far than it is set to a small positive constant. Once S^* is calculated for all cells, we mark an unsuppressed cell, chosen to have a minimal count among those with maximal S^*, for suppression, and re-calculate *S* and S^* for cells in the problem lanes involving the newly-suppressed cell — the other S^* values remain unchanged, saving processing time. We repeat the process until no more problem cells remain, and all cells marked for suppression are suppressed in the released frequency table. Figure 1 illustrates some score calculations for a 3×3×2 table. Darkly-shaded cells are already suppressed, and lightly-shaded ones are in one or more problem lanes. In this iteration, the total for site B and geo 1 will be suppressed, because at 9 respondents, it is the smallest cell among those with the highest score (in parentheses) of 9. After the cell is suppressed, scores will be recalculated for cells in line with the

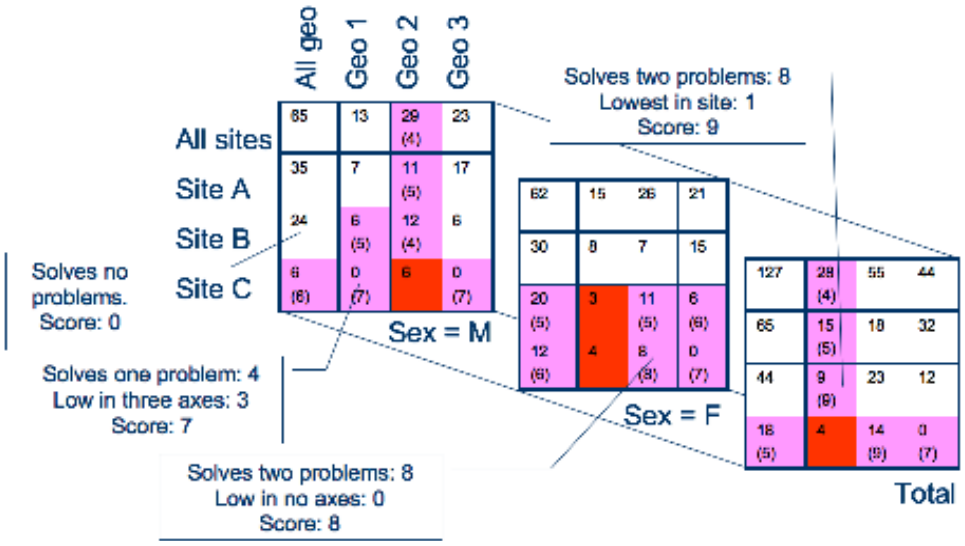

Figure 1 — Example scores for a simple table

suppressed one horizontally or vertically on the "total" table, or in the same position on the tables for male and female.

Like the Merrigan method, this approach suppresses cells deterministically, so long as there is an order to the cells for breaking ties between cells with equally high S^*. Also, the input — a list of table cells with coordinates and counts — is the same as the Merrigan method, so programs required to classify and tabulate cancer incidence were able to be re-used easily. The method is also reasonably fast: even on the larger, four-dimensional tables, the published incidence tables for

all 17 years (1992 to 2008) can be assigned suppression patterns in under 3 hours on a 2×2GHz Xeon desktop machine. Each year included around 500 sensitive cells, and required approximately 250 complementary suppressions on a table with 2268 cells.

Although continuity with older methods and feasibility on available hardware are very useful properties, Tau-Ardent sacrifices a couple of properties relative to CONFID2. First, we make no formal effort to ensure optimality. Output is reviewed by users (subject-matter experts) and accepted on the basis of whether critical cells are published and the tabulations generally look and feel usable. Second, there are cases where disclosure could occur even with no "problems" remaining according to the process. The first drawback is one that the clients accept and work with, and the second drawback is eliminated by a separate verification step based on integer linear programming to guarantee that residual disclosure cannot occur, discussed below.

3.2 Non-square Tables and Special Sites

Non-square tables do not pose a particular problem for the iterative method above. Two types of cells are missing from the CCR Incidence tables: the first type are temporarily missing cells, where a jurisdiction is out of step with other jurisdictions, and the second type are permanently missing "nonsense" cells, which in cancer incidence means cells where there is a mismatch between organ affected by the tumour and sex of the patient, for example prostate cancer in female patients. In both cases, the cell cannot be usefully suppressed: in suppressing a cell, we assume that an attacker can not figure out the value of the cell. If the suppressed value is clearly zero, then we gain nothing from suppressing the cell, and the suppression may not actually protect any sensitive cells that it is assumed to protect.

In the temporarily missing cell case, where a province is missing from the latest data year, not only are that province's cells suppressed, but national totals, missing that province's values, are not published. However, suppression is done as if the province were an impossible combination, in order to avoid releasing cells in the current year that might need to be suppressed in later years when national totals are released and residual disclosure is possible. Currently, we keep no memory of cells suppressed in previous runs, so we trust that tables are similar enough year-to-year that this strategy will suffice. Persistent suppression from release to release is a possibility for future updates to Tau-Ardent.

3.3 Age Clustering

In the Merrigan program, due to the order of searching cells, newborn and toddler cells had a higher likelihood of being suppressed. In many types of cancer, an attacker can reasonably guess that many or all of these cells are zero, due to all the other age groups nearby having zero incidence of that type of cancer. To encourage the system to suppress cells having more uncertainty, a small positive adjustment (enough to break ties with similar cells that are not near non-zero cells) to S^* was made to empty cells adjacent on the age axis to cells with non-zero cases or which are suppressed. This encourages suppressions to cluster with each other and with data on the age axis, which often results in transitions from visible zeroes, to suppressed cells, to cells with many cases. If age has something like a dose effect on probability of a diagnosed case, then this should tend to leave zeroes with low ambiguity visible, while choosing zeroes with more ambiguity for suppression.

3.4 Other fine-tuning

Sometimes it is desirable to adjust the suppression pattern beyond what is easy with score adjustments. In cancer incidence, the subject-matter experts at Statistics Canada are more interested in rates of breast cancer among females than in the both-sexes rate. Despite both sexes being a marginal total, we would rather suppress both sexes cells. An exception can be inserted in the suppression loop which moves the suppression flag from female to both-sexes if the cell chosen by S^* is female. The exception is made before further cells are suppressed because the choice of subsequent suppressions uses previous suppressions. Allowing exceptions in the main suppression loop means that these changes to the suppression pattern are protected by subsequent suppressions as if they were chosen by score.

3.5 Implementation notes

Tau-Ardent is implemented in SAS 9.1 and has been successfully tested in SAS 9.2. It is a set of macros which read settings from user-supplied macro variables, and unweighted count data from SAS datasets. An annotated settings file of SAS statements keeps configuration in one place, and a top-level macro program helps users run all necessary macros in the right order.

4. VERIFICATION AND REPORTING

4.1 Mathematical Programming Solver

While the suppression methodology is heuristic and not provably optimal, the confidentiality verification process on the age-standardized incidence table (Statistics Canada, 2009b) is an exact calculation of the feasibility interval for each cell, as described in Willenborg and de Waal, 1996, pp. 116-119. Two problems are defined for SAS/OR's PROC LP (SAS Institute, 2004) for each cell, one to minimize its value and one to maximize its value, subject to a set of constraints which is constructed once based on marginal totals, non-negativity and integer values for each cell (PROC LP supports integer programming in SAS 9.1).

If the minimum interval is two values and the only requirement is to verify that no suppressed cell can be exactly calculated, then a faster approach is possible. It starts with the construction of a single problem subject to the same constraints as the feasibility problems, with the function to optimize being a constant. The solver is then finished when a feasible solution is found. This initial feasible solution is then compared to the actual values of the suppressed cells, and only cells whose solved values are equal to their actual values are checked with the minimizing and maximizing problems as above, until a solution not equal to the actual value is found. In testing, this resulted in 67 calls (including the initial call and calls to minimize and maximize) to PROC LP for 2003 data, as opposed to 1242 calls using the unfiltered feasibility intervals method above. If the prevention of exact solutions is the level of security required, an initial filtering of cells for feasibility interval verification can save significant amounts of time for large tables or many years of data.

One weakness of the solver (run as described in either of the preceding paragraphs) is that it does not take "inside information" into account in the case of cells whose feasible set includes zero: an attacker who knows that the cell is not empty would have more information than the solver, and if the feasible set included only zero or one cases, then the cell could be calculated exactly. In a future version, the solver should be using this information when trying to calculate the values of individual cells.

4.2 Reports on Suppression Patterns

A major requirement of the new disclosure control system for cancer incidence was extensive output describing the suppression process and output, to complement and reduce reliance on reviewing released tables directly. Both the suppression and solver steps produce summary and detailed information to help assess the suppression pattern and released data.

The suppressor (described in Section 3) produces a report indicating whether the input and output data were consistent with the expected format (valid coordinates, no missing cells, no sensitive cells flagged for release). If any of those checks fail, all cells are flagged for suppression and the year is marked as a failure. After suppression, it produces tables and graphs showing the distribution of cell sizes among complementary suppressions, with information on the overall distribution of cell sizes for comparison, and the distribution of complementary suppressions by cross-tabulation variables. Analysts at Statistics Canada can then spot and investigate any clusters of suppressions which may occur, before releasing tabulations.

The solver (described in Section 4.1) produces a list of suppressed cells, with type (sensitive or complementary) and confirmation that they are not calculable. If feasible intervals are calculated for all suppressed cells, the distribution of these intervals is returned as well as intervals for each cell. The solver output contains confidential data, but is produced in case fine-tuning of suppression is required, and to confirm that the data are safe to release.

5. CONCLUSION

Tau-Ardent is a pragmatic, flexible system for protecting tables of arbitrary dimension with missing or impossible values. Although no explicit efforts or promises are made with regards to minimizing data loss, the system runs quickly on large sets of data, and includes a linear-programming verification step that does guarantee that suppressed values cannot be recovered via the fairly extensive residual disclosure attack described above. The system has been accepted and used on cancer incidence data, but is designed in a sufficiently general way that it should be usable on other frequency data. Future work on Tau-Ardent may address retaining suppression patterns from release to release (protecting against an attacker with every release of the data), improving the solver to deal with more complicated tables or simulate an attacker with more knowledge of a cell, forcing larger feasibility intervals and taking advantage of CONFID2 as much as possible to reduce maintenance and take advantage of its strengths in nested tables and adjustable ambiguity requirements.

REFERENCES

Frovola, O., Fillion, J.-M. and Tambay, J.-L. (2009). "CONFID2: Statistics Canada's new tabular data confidentiality software", Survey Methods Proceedings from the SSC annual meeting, Statistics Canada.

Merrigan, P. (2003). *Program to Suppress Cells with Low Cancer Incidence*, Household Survey Methods Division internal software and documentation.

SAS Institute Inc. (2004). *SAS OnlineDoc® 9.1.3*. Cary, NC: SAS Institute Inc.

Statistics Canada (2009a). *Table 103-0550 - New cases for ICD-O-3 primary sites of cancer (based on the July 2009 CCR tabulation file), by age group and sex, Canada, provinces and territories, annual*, CANSIM (database). http://cansim2.statcan.gc.ca/cgi-win/cnsmcgi.exe?Lang=E&CNSM-Fi=CII/CII_1-eng.htm

Statistics Canada (2009b). *Table 103-0553 - New cases and age-standardized rate for ICD-O-3 primary sites of cancer (based on the July 2009 CCR tabulation file), by sex, Canada, provinces and territories, annual*, CANSIM (database). http://cansim2.statcan.gc.ca/cgi-win/cnsmcgi.exe?Lang=E&CNSM-Fi=CII/CII_1-eng.htm

Statistics Netherlands (2008). τ *ARGUS*. Version 3.3. The Hague: Statistics Netherlands.

Willenborg, L. and de Waal, T. (1996). "Statistical disclosure control in practice". *Lecture Notes in Statistics,* **111**. New York: Springer-Verlag.