Table builder problem - confidentiality for linked tables

Christine M. O'Keefe
CSIRO Mathematical and Information Sciences

Stephen Haslett
Massey University

David Steel
University of Wollongong

Ray Chambers
University of Wollongong

1 Introduction

A central issue when collecting personal and business data is providing the widest range of statistics possible without allowing individuals or firms to be identified. This confidentiality requirement is critical for government statistical agencies internationally, because allowing public disclosure of confidential information, even accidentally, would undermine public confidence in the collection agency and the public's willingness to provide information in future. The consequent drop in response rate for censuses and surveys would in turn undermine the quality of official statistics making them less accurate and useful.

The Australian Bureau of Statistics (ABS) is Australia's official statistical organisation. It is legislatively committed to assisting and encouraging informed decision-making, research and discussion within governments, business and the community. Brian Pink, The Australian Statistician, has said that "We are deeply committed to encouraging access to our statistics." At the same time, the ABS operates within a legislative framework including the requirement to ensure that no person or organisation is likely to be identified, or otherwise put at risk of having their data disclosed.

The ABS would like to offer a 'table builder' service, whereby users can access an internet-based service to specify a statistical table from a variety of surveys that is then delivered to them electronically, without manual intervention or vetting.

The aim of this project was to investigate solutions to the problem of improving access to detailed survey data, while ensuring no person or organisation is likely to be identified, or otherwise put at risk of having their data disclosed, and to link general findings back to the ABS Table Builder problem.

The problem of confidentiality for a single table of data, e.g. average income by age and gender, is an old one, and there is a considerable literature on the topic. For example, see [3, 4, 11]. With more sophisticated computing systems, requests for a particular table are often now dealt with electronically, with automatic checks on confidentiality, and without human intervention in the transmission process. The volume of requests has increased so markedly

that such an automated procedure is imperative. However, when two or more interrelated tables are requested, whether by one user or many, the total information provided can produce confidentiality problems, even if each table on its own does not.

The volume of requests for tabular data from statistical agencies continues to grow, with the implication that it is not enough to confidentialise only single tables, but that it is also vital to ensure that combinations of tables provided cannot be used to identify or derive information about any individual or organisation.

An important aspect of this problem is that the confidentiality method used must be effective against the 'differencing problem'. The differencing problem occurs when a user is supplied with two tables, A and B, that in themselves are both 'confidentialised' and do not disclose any information about an individual. However the user is able to derive a table (A-B) that does disclose such information. In addition, the ABS would like to be able to simultaneously protect the data underlying multiple related tables.

For example, suppose there is a single female aged 90 in a given population with an income in the range $55,000 to $59,999. Regardless of whether there were any older females in the same income range, one table (Table A) could be constructed with maximum age range 90+, income in $5,000 bands and gender, while a second table (Table B) could be constructed using the same income groups and gender but the maximum age range 91+. This would allow the information about the 90 year old's income to be deduced by subtraction or differencing, in contravention of the confidentiality rules. Note that neither Table A nor B breaches confidentiality requirements, but the difference between them does.

There are additional constraints on the desired confidentiality method. For example, it is desirable to maintain additivity and internal consistency of confidentialised tables, so that table entries still add to their marginal totals and apparently different information is not provided in different tables.

While it is essential to find and use methods that ensure confidentiality of provider data is effectively protected, this needs to be done in a way that does minimum damage to the table. While there are several ways in which 'minimum damage' can be interpreted, in practice the priorities could be expressed as:

1. minimise the likelihood of analyses reaching misleading conclusions, due to confidentiality protections, and

2. maximise the likelihood of analyses reaching the same conclusions as they would if carried out on the unprotected data.

The problem is a very broad one, so the MISG participants recommended that, to make progress on this complex and difficult problem in one week, it was best to focus on more restricted instances of the problem. In particular, since the overwhelming majority of table users simply compare cell values, participants focused on this problem, leaving the problem of how to provide access more suited to more sophisticated users conducting statistical analyses in a second phase. In addition, because the table for a business survey may have up to 15,000 responses, it was assumed that a process of averaging over small groups of data points (microaggregation) had been applied to the data that would be tabulated in the Table Builder, using one of the techniques available in the literature.

The participants focussed on making contributions in two main areas, as follows:

1. Identification of sensitive cells in a table

A*B

	A1	A2
B1	7	9
B2	15	3

A*C

	A1	A2
C1	19	6
C2	3	6

B*C

	B1	B2
C1	13	12
C2	3	6

Table 1: Three sub-tables with all cell values at least 3

2. Maximising data utility and minimising information loss - ensuring the table provides useful information.

The outline of the remainder of the paper is as follows: Section 2 focuses on identification of sensitive cells in tables, both for tables containing zeros and tables containing cell counts of three or less. Section 3 provides a very general framework in which the original data are linearly transformed by a stochastic matrix, using a known statistical distribution, to form random linear combinations of the original data under the constraint that certain of the original margins remain unchanged. This framework was found to be a very general one - many current confidentialisation techniques were shown to be examples of this approach. Section 4 considers minimisation of information loss using formal definitions and maximum likelihood techniques, and also outlines other related techniques. Section 5 links these general results back to the Table Builder problem and provides both conclusions and suggestions for further research.

2 Identification of sensitive cells in a table

2.1 Existence of sensitive cells in a table

For there to be sensitive cells in a table, multi-way tables are generally needed. The simplest multi-way table of more than two dimensions has three dimensions and two categories per dimension, ie it is the 2*2*2 three-way table with non-negative count data in each of the eight cells. The surprising thing about even such a simple table is that, in certain circumstances, knowing only certain sub-tables is enough to determine the counts in all the table's eight cells. The required sub-tables are the three, two-way margins, ie if the table is A*B*C, then the sub-tables are A*B, A*C and B*C, with positive count data in all cells except the (1,1,1) and (2,2,2) cells which contain zeros (where the triple (a,b,c) indicates level a for variable A , b for B and c for C). In this situation, the entire table can be deduced from just its sub-tables. This problem is outlined further in [1], page 70, example 3.3.1. It is linked directly to the confidentiality problem because then a user may make multiple requests and get these three two way tables, which individually produce no confidentiality issues, and use them to produce a unique three way table that does.

Consider for example the three sub-tables in Table 1, from [10], page 109, none of which breaches a confidentiality requirement that all table cells equal or exceed three.

[10] show by simple arithmetic, the sum of the cells (A1,B1,C2) and (A2,B2,C1) must be zero, and since counts must be greater or equal to zero both these cells in the underlying 3-way A*B*C table must be zero. Thus, our confidentiality rule that released cells must contain counts of three or more has been breached, at least implicitly. A parallel problem can occur even with continuous data, where there is a known lower threshold (eg zero).

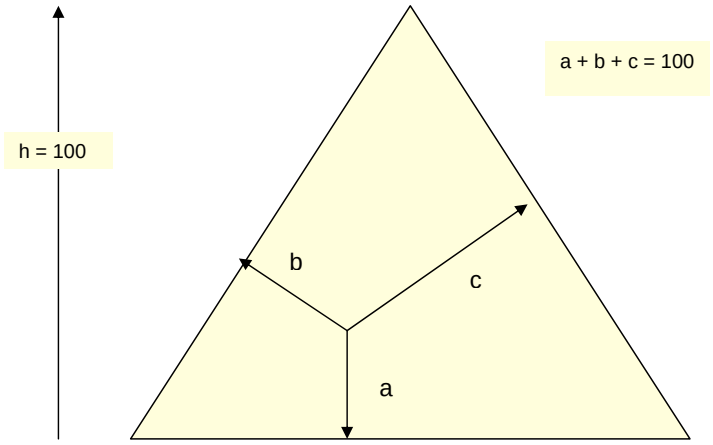

Figure 1: Representation of all cells with value 100 and three contributing individuals

This issue also occurs in tables with more than two categories per variable and more than three variables, which is particularly important for high dimensional multi-way tables because zero cells become increasingly frequent for a fixed sample or population size as the number of cells in the table is increased.

So the next question is how to generalise from 2*2*2 tables to tables which may have more than three dimensions and/or more than three categories per variable. For example in a five way table, if A, B, C, D have 2 categories and E three, the same problem arises if cells (A1,B1,C1,D1,E1), (A2,B2,C2,D2,E2) and (A2,B2,C2,D2, E3) are all zero, and the five possible four way tables are given. The problem can be solved using loglinear models fitted to the complete multiway table, and counting estimable parameters. A more complete discussion is given in the sequence of papers by Fienberg and co-authors. See, for example, [7].

In some cases, problems can exist even when, for example a set of three way sub-tables are requested from a five way table. In this case, even when the requested tables contain no (aggregate) cell with a count under three, there can be zeros in higher (e.g. four way) tables. In this case parameter counting again suffices. See [8] for details.

2.2 An alternative useful structure: the simplex

A method of considering this as a problem on a simplex was discussed and explored. This geometric structure provides a simple visualisation of where confidentiality problems occur, that is, near a boundary of the simplex. A log ratio approach which allows a Hilbert space representation to determining proximity to the boundary was explored.

The method was first proposed by [9] in the case of a cell with three contributing values. As an example, consider a cell in which three individuals contribute values adding up to 100. The collection of all possible values for the three respondents can be represented as points inside an equilateral triangle of height 100, since for any point inside an equilateral triangle; the sum of the distances to the three sides equals the height of the triangle, see Figure 1.

Thus, for example, the top vertex corresponds to the cell of value 100 with contributing values $a = 100$ and $b = c = 0$ while any point on the lower edge corresponds to a cell of value 100 with the contribution $a = 0$. In this diagram, a cell represented by a point on a side of

	A1	
	B1	B2
C1	7	12
C2	2	3

	A2	
	B1	B2
C1	6	2
C2	3	2

Table 2: Example of a 2*2*2 table A*B*C

	A*B	
	A1	A2
B1	9	9
B2	15	5

	A*C	
	A1	A2
C1	19	8
C2	5	6

	B*C	
	B1	B2
C1	13	14
C2	5	6

Table 3: The three sub-tables

the triangle is sensitive, since there are only two (or one) non-zero individual contributions. In fact any cell represented by a point near any of the sides of the triangle is sensitive, as the majority of the value is contributed by only two (or one) individual.

It is a straightforward generalisation to n dimensions for a cell with n contributors. In that case a cell is sensitive if it is represented by a point near a side of the simplex. See [6], Figure 5.

2.3 Confidentiality issues even without zero counts

It is not only being able to deduce all cell counts from sub-tables that is a confidentiality problem for multiple tables. Official Statistics agencies often set a lower threshold, say three, for the count in any cell in a released table. Again requesting multiple tables can allow such a low count to be deduced in a higher dimensional table. Consider again the earlier example, where the zero counts in the (1,1,1) and (2,2,2) cells are now both two. The result is shown in Table 2. If the user now requests the A*B, A*C and B*C tables, they get the result shown in Table 3.

These requests allow the deduction that the sum of cells (1,1,1) and (2,2,2) is four, although the value in either cell cannot be deduced. Since the sum is less than twice the confidentiality limit of three, one of these two cells must be less than three. If this is deemed a confidentiality issue, then setting both these cells to zero in the original table will allow the same method as for zero cells to be used to assess confidentiality risk. Even in this example, it is clear that it is not sufficient to set all sensitive cells to zero and proceed with the checking as before. If cell (1,1,1) were one, and cell (2,2,2) were 4, then the same deduction that one of them is less than three would be possible. What is necessary in higher dimensions is to assess whether sums of specified (blocks of cells) can be less than three times their number, or some other specified threshold. For the 2*2*2 table the four pairs (1,1,1) & (2,2,2), (1,1,2) & (2,2,1), (1,2,1) & (2,1,2), and (1,2,2) & (2,1,1) must be checked to see if their respective sums are less than six.

2.4 Differencing of tables and canonical tables

Consider the simple one-dimensional table in Table 4. Whether the variable in each cell is a

Age (years)	0	1	2	...	100
Count	100	107	93	...	1

Table 4: Table of counts of 100 individuals

	45	46	47	Total
SA	407 (8)	1595 (21)	4003 (24)	6005 (53)
WA	4033 (15)	19695 (25)	10612 (22)	34340 (62)
TAS	**5441 (4)**	**986 (2)**	363 (4)	6790 (10)
NT	**1481 (2)**	**35612 (1)**	0 (0)	**37093 (3)**
ACT	0 (0)	0 (1)	0 (3)	0 (4)
Total	11362 (29)	57888 (50)	14978 (53)	84228 (132)

Table 5: Quarterly Wholesale Sales in $'000

count or some aggregation of a continuous variable by age with counts also given, releasing every one year age bands, 0 to 100, is not possible for this data. Neither is allowing even a single table request from this one way table where any age range of five or more years may be chosen, since then even requesting 0-99 plus 100+ reveals there is only one person over 99. More generally multiple table requests allow deductions about data for finer age ranges than that stipulated in any one table. This is the differencing problem.

Suppressing information is very difficult for multiple table requests because it seems each requested table needs to be considered separately, together with their conjunction. For the one-dimensional table at least, we could specify that groupings of age bands must be the same for all users, and that for example only age 95+ information is available.

But what about higher dimensions? Consider Table 5, which although not real data exhibits the type of problem ABS have. The table is Quarterly Wholesale Sales in $000 for selected States by industrial subdivision. The numbers in brackets are numbers of contributing units. Abbreviations of selected States are SA = South Australia, WA = Western Australia, TAS = Tasmania, NT = Northern Territory, ACT = Australian Capital Territory. Assume cells are sensitive if there are less than three contributors or one contributor accounts for more than 90% of the cell total.

Putting aside cells with zero sales, the cells in bold are sensitive, either because there are less than three contributing units or because one unit contributes more than 90% of the cell total.

Consider Table 6, the corresponding table of counts. Here both marginal tables (State totals and subdivision total) can be released unchanged since all cells are greater than equal to three. However if we want to combine categories so the internal table can be released, then the last two rows must be combined (for example), i.e. making higher dimensional tables confidential requires amalgamations of cells that confidentialise even non-confidential data in collapsed (i.e. marginal) tables, e.g. State.

The idea then that we could produce a canonical table at the finest (i.e. most subdivided) level, and combine cells in this table to form all sub-tables (i.e. all higher dimensional ones, e.g. A*B*C*D to give A*B*C, A*B*D, .. C*D, A, B, C, D) to solve the confidentiality problem, is flawed. What happens is that most collapsed tables and particularly the single

	45	46	47	Total
SA	8	21	24	53
WA	15	25	22	62
TAS	4	2	4	10
NT	2	1	0*	3
ACT	0	1	3	4
Total	29	49	50	128

Table 6: Table of Counts for the Wholesale Sales Data

dimensional ones (A, B, C, D) considerable useful information is suppressed unnecessarily.

The same complications apply when using cell collapsing techniques on a canonical table of tabulated continuous data. The alternative of a different collapsing of every requested table is not practicable or sufficiently secure.

The general conclusion is that to use canonical tables we need to modify data within cells, not combine cells. This idea, which is the topic of Section 3, connects naturally to systems which modify the original microdata (i.e. original records) either in the same way for all records in the same cell, or one at a time before combining them.

Such methods are preferable to cell suppression, which also often requires non-confidential cells to be suppressed, and to amalgamating cells which is rather too dependent on the table or sequence of tables that has been requested.

2.5 Point of control for multiple, linked tables

With multiple, linked table requests by users the risk of disclosure increases. In part, this cannot be controlled, because some users may have supplementary information not contained in the released tables. If users are considered separately, whether each user's requests breach confidentiality can in principle be tested.

But what if users share their tabulated information? Where users know one another (e.g. co-researchers, people working in the same office) they may and likely will exchange information, so it is not possible to maintain confidentiality by simply considering tables released to individuals or even to organisations.

So from the table supplier's (e.g. ABS's) point of view it is necessary that given *all* tables released to *all* users, confidentiality is retained. Otherwise, the table supplier may be held responsible for having publicly released data which in toto breaches confidentiality rules. The boundary determined by all the tables released, considered in concert, may also specify the bounds of a legal responsibility, at least where there is specific legislation governing operation of the statistical agency.

3 A general stochastic framework for confidentialising tables

The discussion in Section 2 raises the question of what other ways are available to confidentialise multiple linked tables. One option is to consider the primary focus not the tables of aggregates but to focus instead on the original unit-level (i.e. individual level) microdata.

We consider using linear transformations of the original data using a stochastic matrix which forms random linear combinations of some of the original data using a known statistical

distribution, under the constraint that certain of the original margins remain unchanged. This framework was found to be a very general one since many current confidentialisation techniques were shown to be examples of this approach.

More formally, consider a set of microdata y and multiplying this by a square matrix A to form y^*, i.e. $y^* = Ay$. Here for non-confidential data, $A = I$ the identity matrix suffices. If averaging, perhaps within subgroups is acceptable, perhaps within subgroups, A will be block diagonal with blocks of size $n_i * n_i$ of the form $n_i^{-1}11^T$ where 1 is a (column) vector of n_i ones. Diagonal forms of A are also possible, e.g. the diagonal elements of A could consist of either 0.9 or 1.1 chosen at random with equal probability for each element of y. Random swapping and permutation of cases are also of this form.

The more general formulation is $A = \alpha_1 I + \alpha_2 S$, where S is a stochastic matrix with for which the row sums are one, and I is the identity. The additional condition $\alpha_2 = 1 - \alpha_1$ may be used. Then when A is multiplied by y it produces what are essentially random linear combinations of the original data. The general extent of the confidentiality is controlled via α , but for any data points that are particularly sensitive there is also additional control via the choice of the corresponding row of A. Further not all elements in each row of A need be non-zero so that those elements of y which contribute to the perturbation of a particular element of y^* and the extent of that contribution can be fine tuned using the choice and size of non-zero elements in the relevant row of A. If a particular element of y is not confidential, choosing the corresponding row of S to have only one nonzero element (which must then necessarily be a one) placed on the diagonal of S ensures that that element in y remains unchanged in y^*.

For the US Bureau of the Census, the cell contribution from each establishment is: (Establishment value)*multiplier+(weight-1) where weight is the sample survey weighting or scaling for unit i in the survey, i.e.

$$y_i^* = y_i(m_i + (w_i - 1)) = m_i y_i + (w_i - 1)y_i \tag{1}$$

where the first term $m_i y_i$ represents the sampled units and the second $(w_i - 1)y_i$ represents the unsampled units. In this situation the sample survey weight w_i offers some protection against disclosing the respondents actual value. Here m_i is random, so for the case where all data is confidentialised and sample size is n. In that case

$$A = \begin{pmatrix} w_{i-1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & w_{n-1} \end{pmatrix} + \begin{pmatrix} m_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & m_n \end{pmatrix} \tag{2}$$

which is of the form $A = \alpha_1 I + \alpha_2 S$.

4 Information loss

The central issue is balancing risk versus utility, or more precisely disclosure risk versus data utility.

4.1 The trade-off between disclosure risk and data utility

In its most basic form, an *R-U confidentiality map* is the set of paired values (R,U) of disclosure risk R and data utility U that correspond to various strategies for data release. Figure 2 shows

hypothetical values of disclosure risk and data utility for the three strategies of releasing no data, an original dataset and for the confidentialised version of that dataset.

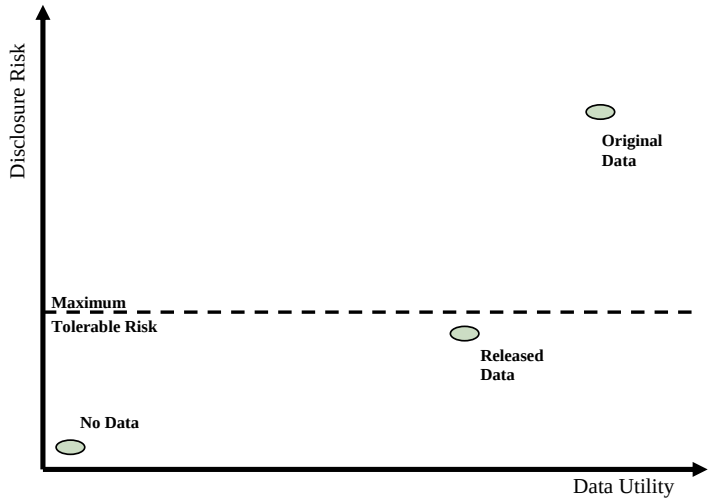

Figure 2: An R-U confidentiality map for three data sets

Often an R-U map is drawn for a single strategy which involves the implementation of a disclosure control technique with a choice of parameters, for example, the addition of noise which depends on the magnitude of the error variance. As the parameters vary, a curve is mapped in the R-U plane, as is shown for a hypothetical example in Figure 2. This curve portrays the trade-off between disclosure risk and data utility, enabling an informed decision to be made. To date examples of R-U confidentiality maps have been constructed for some disclosure control techniques (for example, the addition of noise) and certain intruder objectives. See [5].

4.2 Utility measures

Utility measures can include distortion of distributions of random variables, variance of estimates, impact on measures of association, impact on goodness of fit, overlap of confidence intervals, cells suppressed and values suppressed. As an alternative, we investigate the use of information loss as a measure of utility using formal definitions and maximum likelihood techniques.

4.3 Formal structure of information loss

Consider the situation in which the original data is d_u and this is reduced by applying a statistical disclosure control technique to give d_R. Assume that an analyst is making inferences using Maximum Likelihood Estimation based on some parametric model. Based on d_u we have the score function, $\mathrm{SC}u(\theta; d_u)$ and observed information matrix $\mathrm{info}_u(\theta; d_u)$.

For this formulation, let $d = \{d_u, d_R\}$, then applying the missing information principle (see [2]) gives

$$\mathrm{SC}R(\theta; d_R) = E[\mathrm{SC}u(\theta; d) \mid d_R] \neq \mathrm{SC}u(\theta; d_R) \quad (3)$$

$$\mathrm{info}_R(\theta; d_R) = E[\mathrm{info}_u(\theta; d) \mid d_R] - V(\mathrm{SC}u(\theta; d_R)) \neq \mathrm{info}_u(\theta; d_R). \quad (4)$$

Assume now that statistical disclosure control techniques are applied, such that the density $f(d_R \mid d_u; \phi)$ does not depend on θ and ϕ is known. The likelihood is

$$\ell(\theta; d) = \log(f(d_u; \theta) + \log f(d_R \mid d_u; \phi) \tag{5}$$

$$\text{so that } \frac{\partial \ell(\theta; d)}{\partial \theta} = \text{SC}u(\theta; d_u) + 0 \tag{6}$$

$$-\frac{\partial^2 \ell(\theta; d)}{\partial \theta \partial \theta_T} = \text{info}_u(\theta; d_u) \tag{7}$$

Hence,

$$\text{SC}_R(\theta; d_R) = E[\text{SC}_u(\theta; d_u) \mid d_R] \tag{8}$$

$$\text{info}_R(\theta; d_R) = E[\text{info}_u(\theta; d_u) \mid d_R] - V(\text{SC}_u(\theta; d_u) \mid d_R) \tag{9}$$

which explicitly shows the loss of information.

As an application, suppose that the released data (with statistical disclosure techniques applied) are

$$y_i^* = \gamma_i y_i \qquad x_{k_i}^* = \gamma_i x_{k_i} \tag{10}$$

where γ_i are independent random variables. The data vectors are $y^* = \Gamma y$ and $x^* = \Gamma x$ so that $y = \Gamma^{-1} y^*$ and $x = \Gamma^{-1} x^*$ where $\Gamma = \text{diag}(\gamma_i)$. For linear regression, the score based on d_u is

$$X'(y - X\beta) = X^{*\prime}\Gamma^{-1}(\Gamma^{-1}y^* - \Gamma^{-1}X^*\beta) = X^{*\prime}\Gamma^{-2}y^* - X^{*\prime}\Gamma^{-2}X^*\beta. \tag{11}$$

The score based on $d_R = (y^*, X^*)$ is the expectation of this given d_u, namely

$$X^{*\prime}E[\Gamma^{-2}]y^* - X^{*\prime}E[\Gamma^{-2}]X^*\beta \tag{12}$$

Thus, we end up weighting by $E[\gamma_i^{-2}]$. Note that the distribution of γ_i must be such that Γ^{-1} and $E[\Gamma^{-1}]$ exist, and that commonly $E(\gamma_i) = 1$. In the case of aggregation, $y^* = A'y$ and $X^* = A'X$ where A is an aggregation matrix. We note that the direct approach may be easier for linear models, where $y \mid X \sim (X\beta, \Sigma)$ implies that $y^* \mid X^* \sim (X^*\beta, A'\Sigma A)$ for A constant.

The conclusion is that a formal approach to information loss is possible, so we can assess the effect of changing even multivariate data. Risk is however more difficult to quantify because it has a subjective element.

5 Conclusions

We recognise that the confidentiality problem is very complex and requires much more extensive research for a comprehensive solution. This view is supported by the fact that it is an active research area internationally, with no clear consensus on the best approach to confidentialising tables. Our conclusion is that it would be desirable for someone to immerse themselves in the problem and to examine the trade-off between data utility and information loss for real ABS survey data.

In summary, progress at MISG has provided a framework for the overall process which included settling on the approach of perturbing the underlying microdata rather than perturbing the table cells. In addition, the participants made contributions in identifying sensitive cells and in determining the consequential information loss for a given class of microdata perturbation methods.

Acknowledgements:

The authors thank Robert Mellor for assisting with the moderating and Wilford Molefe for participating as the student moderator on this problem.

References

[1] Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. (1975) *Discrete multivariate analysis: theory and practice*, MIT Press, Cambridge.

[2] Breckling J.H., Chambers R.L., Dorfman A.H., Tan S.M. & Welsh A.H. (1994) Maximum likelihood inference from sample survey data, *Int. Stat. Rev.* **62**, 349-363.

[3] Domingo-Ferrer J. & Torra V. (Eds.) (2004) Privacy in Statistical Databases, *Lecture Notes in Computer Science,* **3050**, Springer-Verlag Berlin Heidelberg.

[4] Doyle P., Lane J.I., Theeuwes J.J.M. & Zayatz L. (2001) *Confidentiality, disclosure and data access: theory and practical applications for statistical agencies*, Elsevier, Amsterdam.

[5] Duncan, G.T., Keller-McNulty, S.A. & Stokes, L. (2001) Disclosure Risk vs Data Utility: The R-U Confidentiality Map, *Los Alamos National Laboratory Technical Report* LA-UR-01-6428. 2001.

[6] Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G. & Barcelo-Vidal, C. (2003) Isometric Logratio Transformations for Compositional Data Analysis, *Math. Geol.* **35**, 279-300.

[7] Eriksson, N., Fienberg, S.E., Rinaldo, A. & Sullivant, S. (2006) Polyhedral conditions for the nonexistence of the mle for hierarchical log-linear models, *J. Symb. Comp.* **41**, 222-233.

[8] Haslett, S. (1990) Degrees of freedom and parameter estimability in hierarchical models for sparse complete contingency tables, *Comp. Stat. Data Anal.* **9**, 179-195.

[9] Robertson, D.A. & Ethier, R. (2002) Cell suppression: experience and theory. Domingo-Ferrer, J. (Ed.) *Inference control in statistical databases: From theory to practice* State-of-the-Art Survey, LNCS 2316. Springer-Verlag Berlin Heidelberg.

[10] Willenborg, L. & de Waal, T. (1996) Statistical disclosure in practice, *Lecture Notes in Statistics* **111**, Springer, New York.

[11] Willenborg, L. & de Waal T. (2001) Elements of Statistical Disclosure Control, *Lecture Notes in Statistics* **155**, Springer, New York.