

PROTECTING THE CONFIDENTIALITY OF STATISTICAL TABLES PRODUCED BY AN AUTOMATED "TABLE-BUILDER"

Overview of problem

The aim of this project is to improve access to detailed Australian Bureau of Statistics (ABS) survey data, while maintaining a legislative requirement to ensure no person or organisation is likely to be identified, or otherwise put at risk of having their data disclosed.

The problem is particularly acute for business data - as there can be real commercial advantages to obtaining the data of a business competitor. However, it is just as important to ensure household data are protected. In order that the ABS maintain the trust of the community, the methods we use must withstand scrutiny and be publicly defensible - it must be clear that adequate protection is provided and that the risk of identification or disclosure of information for any particular person or organisation is very low.

However, the ABS also has an obligation to ensure that the Australian community gets the maximum benefit from the information collected by the ABS - which means the ABS needs to provide a responsive and flexible statistical output service.

In particular, the ABS would like to offer a 'table builder' service, whereby users can use an internet-based service to specify a table that is then delivered to them electronically, without manual intervention or vetting. This means that the tables that are delivered must be automatically 'confidentialised' to ensure that no table or combination of tables can be used to identify or derive information in respect of any individual or organisation.

While it is essential to ensure that the confidentiality of provider data is effectively protected, it is also highly desirable to protect the tables in the way that does minimum damage to the data. While there are several ways in which 'minimum damage to the data' can be interpreted, in practice the priorities could be expressed as: (i) minimise the likelihood of analyses reaching misleading conclusions, due to confidentiality protections, and (ii) maximise the likelihood of analyses reaching the same conclusions as they would if run on the unprotected data.

An important aspect to 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. However the user is able to derive a table (A-B) that does disclose information. For example, if table A reports on all people in a geographic region aged 70-80 and table B reports all people in the same region aged 70-79. By differencing they can obtain a table for all people aged 80 that may, for example, be based on the responses of a single individual. In general, any linear combination or union of 'safe' tables produced by the method must also be 'safe' and not allow the identification of any individual or organisation.

What makes a table 'unsafe' to release

There are two key risks to inappropriate disclosure that must be addressed in any release of statistical tables.

Firstly, cells with very small numbers of contributors. The danger occurs when a table user knows that there is only one person in the population with a particular set of characteristics, or has very good knowledge of a small number of units with a particular set of characteristics. For example, if a user knows there is only one person aged 80 years in a particular area, then a table of average income for 80 year olds in the area would identify the individual in question and disclose their income. Or if a user was aware there were only two large legal firms in a particular area and they knew the income of one of these firms, then a table of average income would disclose the income of the other firm. (A example is where the user is one of the two firms, and they are using the table to get information on the performance of a competitor). These two example illustrate why cells with only one or two contributors are of particular concern. However, there is also some risk attached with other small contributor cells as well (eg three businesses in a cell, where a knowledgeable user knows or can closely guess the income of two of these businesses).

The second key risk is where one or two units heavily dominate a cell. For example, if there are a number of manufacturing businesses in an area, with one accounting for 95% of all income earned by manufacturers in the area. Because this one unit so heavily dominate the cell, the cell total will be very close to the individual business income of this dominating unit, and change in the cell over time will closely correspond with changes in the income of this one dominating business. The dominance problem is sometimes expressed via "the p% rule" of confidentiality - which states that no user should be able to derive a characteristic of any individual respondent to within p%. For example, no user should be able to determine the income of the manufacturing business to within + or - 20% of its true value. In this example, if the individual business income of the largest contributor is \$100m, then any cell total between \$80m and \$120m would be considered 'unsafe'.

The ABS has an obligation to protect the confidentiality of all information directly provided or derivable from information directly provided by respondents. This includes not just individual responses, but derived items (eg business operating profit is derived from components of income and expenditure) and movements in values over time for repeating surveys (eg monthly change in sales).

Methods of confidentialising statistical tables

Cell suppression

One widely used method that is generally considered to minimise damage to the data in a table is the method of cell suppression. For example, suppose the cell shown in red below is 'unsafe' due to a small number of contributors.

Age	Region A	Region B	Region C	Region D	Total
0-14	3	4	6	5	18
15-24	4	9	15	12	40
25-54	5	12	20	15	52
55+	1	3	5	4	13
Total	13	28	46	36	123

The cell can be protected by suppressing its value. Secondary suppressions must also be made to prevent the unsafe value being derived from the table marginals.

Age	Region A	Region B	Region C	Region D	Total
0-14	3	4	6	5	18
15-24	4	9	15	12	40
25-54	*	*	20	15	52
55+	*	*	5	4	13
Total	13	28	46	36	123

While cell suppression is effective for a single table, or a small number of pre-specified tables, it is difficult to manage in the context where further ad hoc tables can be released. For example, if a user now requests the same table but with region A excluded:

Age	Region B	Region C	Region D	Total
0-14	4	6	5	15
15-24	9	15	12	36
25-54	12	20	15	47
55+	3	5	4	12
Total	28	46	36	110

They obtain the value of the secondary suppression cells in region B and can use these to derive the missing cell values in region A. Or if this is prevented by recognising and treating the unsafe cell values in region B, as in the table below, then more secondary suppression cells are created. It becomes increasingly difficult to manage suppressed cells as more and more tables are created.

Age	Region B	Region C	Region D	Total
0-14	4	6	5	15
15-24	9	15	12	36
25-54	*	20	15	*
55+	*	5	4	*
Total	28	46	36	110

Although the problem seems very difficult, there is work in the US looking at a

system for determining how to apply suppressions so as to maximise the supply of useful information that can be provided in future tables.

Cell perturbation

An alternative to cell suppression is cell perturbation. One of the simplest forms of cell perturbation is random rounding to base k . The most commonly used values for the rounding base are 3 and 5. If base 3 is used, then any multiple of 3 remains unperturbed. If the cell value is $1 \pmod 3$ then the cell value is rounded down to the nearest multiple of 3 with probability $2/3$ and rounded up to the next multiple of 3 with probability $1/3$. If the cell value is $2 \pmod 3$ then it is rounded down with probability $1/3$ and rounded up with probability $2/3$. These probabilities are chosen to give zero expectation on the final perturbation.

So, for example, the table of counts by age and region considered above may be perturbed as follows. In this first table all cells have been independently rounded, and as a result the table is no longer additive.

Age	Region A	Region B	Region C	Region D	Total
0-14	3	3	6	3	18
15-24	3	9	15	12	39
25-54	3	12	18	15	51
55+	0	3	6	6	15
Total	15	30	48	36	123

In this second example, additivity is maintained by re-deriving marginals as the sum of the perturbed interior cells:

Age	Region A	Region B	Region C	Region D	Total
0-14	3	3	6	3	15
15-24	3	9	15	12	39
25-54	3	12	18	15	48
55+	0	3	6	6	15
Total	9	27	45	36	117

Although this method aggregates cell perturbations to the marginals, putting a relatively large amount of noise (i.e. large variance of the final perturbation distribution) into the marginals. An alternative is to try to maintain the values of the marginals and restore additivity by adjusting interior cells eg as in the example below:

Age	Region A	Region B	Region C	Region D	Total
0-14	3	3	9	3	18
15-24	3	9	15	12	39
25-54	6	12	18	15	51
55+	0	3	6	6	15
Total	12	27	48	36	123

We can generalise the perturbation approach by considering generic cell perturbations e that have zero mean. We are then free to play around with the shape of the distribution and its variance to give the best balance between disclosure risk and information loss in the table (i.e. damage done to analytical outcomes).

Cell perturbation seems attractive when differencing can occur - as differencing two tables results in a difference that is itself affected by perturbation, and so 'safe' (so long as we can be assured that final perturbation on the difference provides sufficient protection).

One potential weakness of an unbiased perturbation approach is that a user may be able to determine true cell values by requesting a large number of realisations of the cell value. For example, if a user requests a table of region A and region B to obtain one realisation of region A values, then requests a table of region A and region C to get a 2nd realisation, then region A and region D, and so on. If independent perturbations are applied each time, then by averaging over all realisations the user can obtain the original value. This danger can be overcome by ensuring that any cell value always receives the same stochastic outcome from a perturbation distribution. One method of achieving this is to assign record keys to each individual record of data. Combining records to produce a cell value then generates a cell key that is produced by combining the record keys of the contributing records. This cell key is then used to control the point in the perturbation distribution that the perturbation realisation is selected from.

Development of a table-builder confidentiality method to date

The attached paper discusses some perturbative approaches that have been considered in the context of the table-builder problem - particularly in the context of business survey tables where the dominance problem must be considered as well as the small number of contributors problem.

The EZS method has some very attractive properties - in particular the way it will apply a sufficient amount of protective noise to cells that are heavily dominated by a single unit, with the noise generally 'cancelling out' for cells with a large number of contributors and no units dominating. However it is vulnerable, as described in the attached paper, to users deriving the value of the noise factor applied to an individual unit and undoing the protection. The EZS method may well make a good base for a final method, if we can design something to address the vulnerabilities.

The attached paper - and this briefing note - describe the generic problem and seem to be suggesting a single solution can be found to address any sort of statistical table. However, this need not be the case, and it is likely that the best overall result will be obtained by considering different types of statistical tables quite distinctly. Accordingly, we suggest that the following are the broad classes of statistical tables that need to be supported by the table-builder.

"Household surveys" - or others with small sampling fractions and categorical data items

The standard sampling model used for ABS household surveys applies a common probability of selection to each person within a State/territory. This means all units have low probability of selection (in the order of 1 in 100 or lower) and some protection is provided by the sampling. Estimates are performed by weighting and aggregating survey records, so for example a cell with a single contributor may have a cell value in the order of 100 reflecting the estimation weight attached to the single contributing record. Household surveys also deal almost exclusively with categorical data, and so table cells represent weighted counts of persons or households with a particular combination of characteristics. The main example of non-categorical data in household surveys is income - and of course it is possible to convert the small number of non-categorical variables into categorical variables. (Note - variables like age or number of persons in the household take a finite number of integer values and can be considered categorical).

"Quantitative business surveys" - or others with high sampling fractions and magnitude variables

The second class of surveys covers ABS surveys such as Monthly Retail Sales, Quarterly Business Indicators, Quarterly Capital Expenditure and the like. Typically they are focussed on a small number of quantitative (\$ value) variables, such as sales, wages, profit, capital expenditure and the like. Large businesses have a large impact on aggregates, and so the largest businesses will be selected with certainty for these surveys. Therefore there is no protection provided by sampling for these large businesses - a data user knows that these large businesses are in the survey data file somewhere which heightens their risk of identification. The dominance problem is the cause of most confidentiality issues. The quantitative data will be supplemented with a typically small number of categorical variables that usually are restricted to business demographics eg industry, state/territory, size, public/private.

"Quantitative and descriptive business surveys" - or other with high sampling fractions and a combination of categorical and magnitude variables

The final category of survey data is typified by annual industry-focussed surveys, which collect detailed financial data (and so have a large number of quantitative variables with additive and other relationships) as well as categorical information that extends beyond simple demographics. For example, an annual manufacturing industry survey may collect detailed income and expense items, and also categorical variables that may indicate whether the business was an exporter, whether or not it produced certain types of commodities, whether or not it incurred certain types of expenses, characteristics of innovation, and so forth. It is exposed to the cell dominance problem and also the 'small number of contributors' problem where a business may be identifiable through a unique combination of categorical variables. Certain units will be selected with certainty and so afforded no protection from sampling. This class will be the class that is hardest to protect effectively, and consequently the class where we would expect to see the greatest amount of information loss in tables (or distortion of analytical outcomes).