WP.22
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Manchester, United Kingdom, 17-19 December 2007)

Topic (ii): Tabular data protection

CENSUS TABLES: UTILITY AND SAFETY VIA A CELL THRESHOLD

Supporting Paper

Prepared by Mike Camden, Paul Cowie and Lisa Henley (Statistics New Zealand)

Census tables: utility and safety via a cell threshold

Mike Camden*, Paul Cowie* and Lisa Henley*

*Statistics New Zealand, PO Box 2922, Wellington, New Zealand
mike.camden@stats.govt.nz, paul.cowie@stats.govt.nz, lisa.henley@stats.govt.nz

Abstract: Tables of counts from a population census are valued highly by users, especially users from local government. Users request detailed tables that are sometimes very sparse. Statistics New Zealand's past protections against sparseness include random rounding to base 3 and a mean cell size rule. It decided, for its 2006 Census of Population and Dwellings, to further enhance data utility and protect safety by setting a threshold for sparse tables: counts above this are released and counts at this value or below are suppressed. The confidentiality rules are applied to each geographical area separately, and our current geographical areas vary widely in population. This variation in size is a source of sparseness, but we use it in two ways. We calculate measures for utility and safety for each geographical area, and use them to evaluate the threshold approach. We also use them to help in setting the size for a possible new set of geographical areas designed for output. The results can be applied to heighten utility and safety for our 2011 Census.

1 Introduction: sparseness and methods of managing it

The New Zealand Census of Population and Dwellings is held every five years. The most recent one was in March 2006, and planning is in progress for the 2011 Census. Tables of counts are a major mode for dissemination of census information. These tables are often sparse, with large proportions of counts being 0's and 1's. The paths for providing confidentiality protection for tables like these are summarised by Wooton and Fraser (2005) and Schlomo (2005). Statistics New Zealand has taken one path: random rounding, a mean cell size rule for tables by geographical area, and now a threshold rule for cells. We refer to these rules as R, M and T, respectively, and to the combinations of interest to us as R, RM, RT and RMT. In choosing our path, acceptance by users is important.

We use the utility/risk framework (Duncan, Fienberg, Krishnan, Padman and Roehrig, 2000) to assess our set of rules. We refer to the converse of risk as safety. We calculate some measures for utility and safety for each geographical area, and use the variation in the size of the areas to assess the effect of our rules. We use this variation also to indicate the ideal size for geographical areas designed for output.

1.1 Recent history of confidentiality for the Census of Population and Dwellings

Counts in tables have been protected since the 1981 Census by random rounding to base 3 (R), and since 2001 by a mean cell size rule (M). The mean cell size for 2001 was 1 unit per cell. The 2001 rules allowed the release of sparse tables containing

many 0's and a few 3's, which are likely to arise from 1's. Jackson, Corscadden and Zeng (2005) examined uniqueness in the 2001 Census, and recommended a design perspective for small geographic areas and categories with small proportions. In preparation for 2006, a modified set of rules was created that targeted sparseness by raising the mean cell size from 1 to 2, and applied M to each geographic area separately. (The table for an area was published if mean cell size > 2.) It targeted the sensitive variable income by using an aggregated version without small frequencies.

This set of rules was designed so that, when clients submitted a request, staff could tell them whether counts would be supplied, and for which areas. The rules used information from outside the table (population in geographical area, number of cells), and so could be applied before tables were built. The rules worked well for standard tables and large geographies. As client requests for customised tables from the 2006 Census flowed in, it became clear that clients had ongoing needs for tables that contained both sparse patches and useful counts, and failed M. We needed to provide both safety and utility for these customised tables. Our solution was to keep the mean cell size rule M, but where the table for an area failed it, we would release counts above the threshold of 5, and suppress counts of 5 and below (T). We continued with random rounding (R), so secondary suppression was not needed. We assume that rules RMT reduce risk to an acceptable level.

In moving along this path, we shifted from rules that could be applied before tables were built to further modifications that would be applied after tables were built. We responded to a complex set of pressures with a complex and more finely targeted solution. This path raises questions about both utility and safety. The paper quantifies the effect of the path on utility and safety, across a range of geographical sizes.

1.2 Variation in geographical classifications: risks and opportunities

Our rule set recognises the importance of geographical location in confidentiality protection, and treats the table for any geographical area as a separate entity. The rules are applied area by area. The main geographical classifications are these:

1. 1 country, with a usually resident population (2006 Census) of 4,249,737 people
2. 16 regional authorities: min = 31,326 mean = 251,708 max = 1,303,068
3. 73 territorial authorities: min = 612 mean = 55,172 max = 404,658
4. 1,918 area units: min = 0 mean = 2,100 max = 9,027
5. 41,384 meshblocks: min = 0 mean = 97 max = 1,431.

We are concerned here with the last two classifications. Both are extremely variable, and were designed for purposes like data collection and classification-building rather than for dissemination. Fig 1 shows the size distribution for area units. The large number of areas with low sizes, on the left, is a major source of sparseness. However, we make use of this variation to assess the effects of our rules on utility and safety, and to suggest a better-distributed new set of geographies for the 2011 Census.

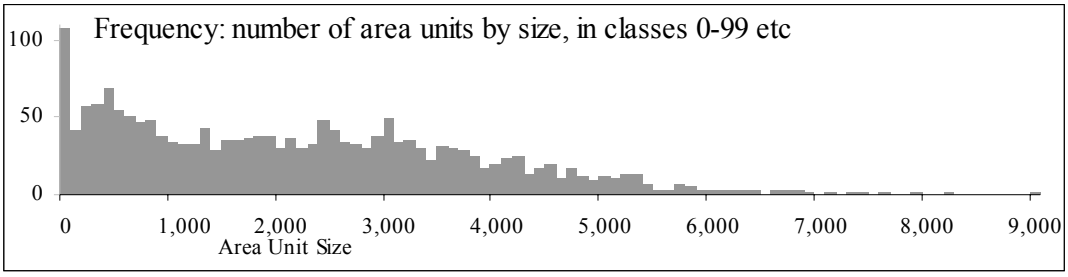

Fig 1 Frequency distribution of Area Unit Size (in classes 0–99, 100–199 etc).

1.3 Aims and contents of this paper

This paper aims to provide information about two questions. The first is about how well the path that we are taking provides utility and safety. The second is about how new classifications, especially for geographic location, can be designed. Part 2 uses a one-way table to illustrate how our rules work, and introduces the two-way table used in the analysis. It discusses table structure and a table's frequency distribution of counts. Part 3 defines some measures, applies them to our typical table and seeks answers to our two questions. It checks for new risks from a complex rule set. Part 4 concludes, with messages from the data and possibilities for the future.

1.4 Example 1: a one-way table

Table 1 concerns a rural area unit that is large in area but small in population size. Its subject population, for the 10-category variable Occupation, is 19.

Occupation:	a	b	c	d	e	f	g	h	i	j	Tot	SD
Row 1: Raw	16	0	0	1	0	0	1	0	0	1	19	–
2: R	15	0	0	3	0	0	0	0	0	0	18	–
3: noise	-1.0	0.0	0.0	2.0	0.0	0.0	-1.0	0.0	0.0	-1.0	–	0.8
4: RM	c	c	c	c	c	c	c	c	c	c	18	–
5: noise	-14.2	1.8	1.8	0.8	1.8	1.8	0.8	1.8	1.8	0.8	–	4.7
6: RMT	15	c	c	c	c	c	c	c	c	c	18	–
7: noise	-1.0	-0.3	-0.3	0.7	-0.3	-0.3	0.7	-0.3	-0.3	0.7	–	0.5

Table 1 Counts for the 10 categories of Occupation, for a small area unit with 19 people, with three rule sets and the noise resulting from each; c = confidential.

Row 1 has the original counts. It can be reduced to its proportional frequency distribution: $P(0) = .6$, $P(1) = .3$, $P(16) = .1$. The mean cell size is 19/10 = 1.9.

Row 2 has random-rounded counts (R), and row 3 has the noise added by this process. The table passes M with parameter 1, and would have been published for 2001. The standard deviation of the noise, SD, measures information disturbance.

In row 4, we use rules RM with parameter 2 for M (as in our 2006 rules), and the table fails. A user could estimate the c's simply as the apparent mean: 18/10 = 1.8. The noise that results is in row 5, and has a higher value for SD.

Row 6 represents our current practice: rules RMT. The table fails M, the threshold of 5 is applied, and the one count over 5 (the 15) is rescued (rule T). Rule R is applied. A user could replace the c's with an improved guess: 3/9 = 0.33. The noise that results is in row 6, and has a lower value for SD.

1.5 Example 2: a two-way table, and the structures within it

Our analysis below uses a typical user request for customised output. For each area unit, we find the two-way table of counts for:
Travel: main means of travel to work: 11 categories: foot, cycle, bus, train, car etc
Age: age in 5-year groups: 11 categories: 15–19, 20–24, … to 65+.
There are 121 cells, and the limit for rule M is 2 x 121 = 242 people.

The count for any cell depends on some structural elements: the cell's values for Travel and Age, the interaction between these, and the adjacent values for Age. These structures result in the actual counts, and we can convert these counts into their frequency distribution. This is a rich source of information about utility and safety, and we use it to define measures of both.

If we assume nothing about a table apart from population N and number of cells K, then we could assume the counts are approximately Poisson, with parameter = N/K. Some of our graphs include curves for this model. They show that real tables have much more clustering than tables from the model would have.

2 Measuring the effect of our rules on utility and risk

We define one measure for safety and two for utility, view them against Area Unit Size for area units of size 1 to 3,000, and assess risks arising from the complexity of our RMT rule set.

2.1 Three measures defined

Let $P(x)$ = the proportion of cells with count x. $P(0)$, $P(1)$, $P(2)$, $P(3)$ and combinations of them are all useful as measures of safety (or its converse, risk). All these proportions, and the relationships among them, can be investigated using our set of area units. Our measure is: PS = $P(0) + P(1)$, the Proportion of Sparse cells. Both 0's and 1's are high-risk. The 2's are also risky, but are not used here. PS measures risk in the original table, before any rules are applied. Our rules reduce this risk: R replaces 0's, 1's and 2's with 0's or 3's, and M with T replaces them with c's for small area units.

Our first measure of utility is PR: the Proportion of people for whom (rounded) counts would be Released, assuming that they needed to be Rescued by the threshold (T) from suppression. PR measures utility only where T is used.

Our second measure of utility (or in fact a converse of it: information disturbance), SD, arose earlier. It is the standard deviation of the noise added by a set of rules. It compares the counts from any rule set with the original counts. We calculate and graph it for these four rule sets: R, RM, RT, RMT. It is related to variance (Wooton and Fraser, 2005) and mean absolute deviation (Schlomo and Young, 2005). SD has the same units as the counts that we are interested in, and is easily calculated.

2.2 Some data visualisation from our two-way table

The two-way table from Example 2 in Part 1.5, for Travel by Age, has 121 cells, and can be produced for each of our area units. The graphs below show the measures for those area units with subject population from 1 to 3,000. The areas above 3,000 have low risk and minimal loss from the rules, and so are omitted.

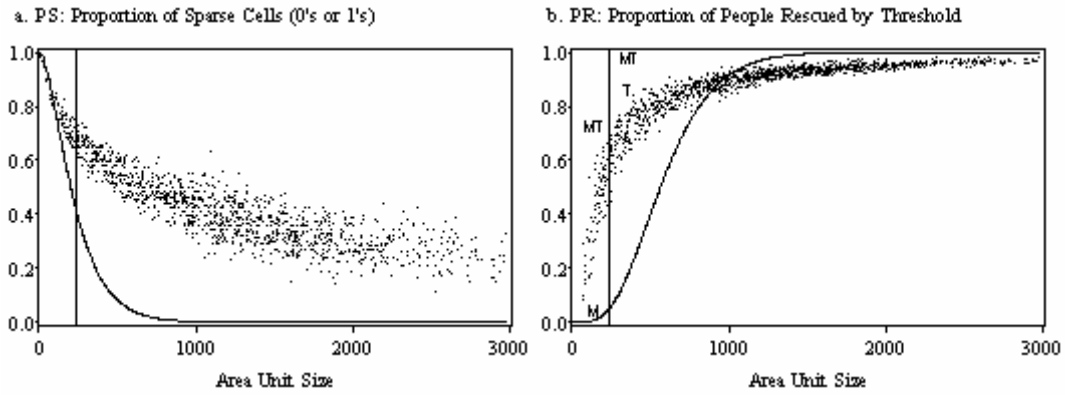

Fig 2: a: PS: Proportion of Sparse cells containing 0 or 1, by size of area unit. b: PR: Proportion of people Rescued from suppression by T, assuming that their area failed M and therefore needed rescue. Both graphs contain curves for the Poisson model.

Fig 2a shows risk, measured by PS = $P(0) + P(1)$. It is high for small areas, and drops off as area size increases. Small areas need protection, and the graph throws some light on what 'small' means here. For this table, our rule M (with parameter 2) protects areas up to 121 x 2 = 242. The graph suggests that the parameter 2 is not excessive. Protection could, in future, come from a new geographical classification. Graphs of further examples will suggest where the minimum size for this could be. (Much of the downward trend in this graph comes from $P(0)$; $P(1)$ is quite stable.)

Fig 2b shows utility, measured by PR. It helps to answer the question as to whether rule T, rescue by threshold, is worthwhile. The vertical line is at the limit for M: size = 242. To the left of the line are the areas that fail M. When we include rule T, the proportion of people whose data are released rises from zero (see M on the graph) to

the point cloud (see MT). Fig 2b also helps with the question as to whether we now need the mean cell size rule at all: we could suppress all counts below threshold, regardless of area size. Our rule set would be RT only. To the right of the line, under RMT, all areas pass M, and all counts are released (see MT). Under RT, the proportion of people whose data are released would fall to the point cloud (see T). The graph suggests that RMT is worth keeping.

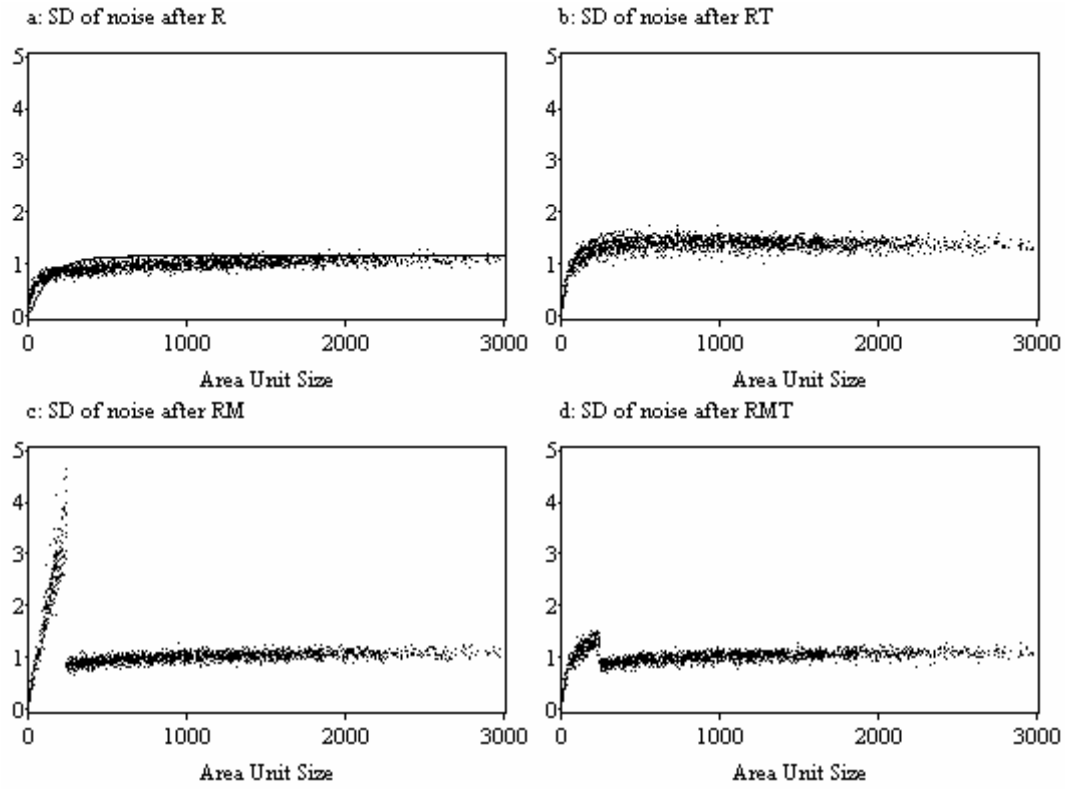

Fig 3 SD: Noise after rules R (with Poisson), RT, RM, RMT; by Area Unit Size.

Fig 3a shows that, under random rounding only (R), the noise rises up from 0 (zero cells stay zero and have no noise added) to a constant level. If a table has no 0's and counts spread evenly, we expect this level to be the noise for R: $2/\sqrt{3} = 1.155$, but it is lower. We have added the curve for Poisson tables, which does reach this level.

Fig 3b shows that, with rules RT, the noise is consistently higher. This helps further with the question as to whether we now need the mean cell size rule at all. Under RT, we would suppress cells below the threshold for all areas, large and small. We would eliminate all sparseness, but pay for it with the increased noise.

Fig 3c shows what happens with our earlier plan to use rules RM. The noise rises steeply as Area Unit Size approaches the limit for M.

Fig 3d shows the improvement in moving to our current plan, RMT. This helps with the question as to whether the rescue by threshold is worthwhile. The graph implies that it is: information disturbance is smaller and so utility is improved.

The parameters for R (base = 3), M (mean cell size = 2) and T (threshold = 5) could all be varied, and the effects assessed with these measures. Our threshold is set at 5 for several reasons. Numbers like 2, 5, 8 … as thresholds have a much tidier interaction with R than the others, with less bias and less disclosure risk. The threshold at 5 hides 0's, 1's and 2's as it must, and goes a little further. Users lose counts rounded to 3, but these have a high proportion of noise added by R. The proportion of people lost by T is quantifiable, and appears as 1 - PR in Fig 2b.

We have used these measures on other tables with area units, and on tables with the meshblock classification. The patterns and conclusions are similar. We need to extend the measures to test counts of households.

2.3 Complexity of the rule set

The use of the complex RMT combination of rules raises three concerns. The first is about complexity for users. Our discussions with expert users of tables suggest that they are not concerned about this, and that they see our path as a logical and useful one. The second is about complexity for automation of the rules. The rules have been programmed successfully into our software systems. The third is about interactions among the rules, and the effect on disclosure control. There are combinations of counts for which the rules provide a lower level of protection than usual. We balance this risk against the reduction in published 3's, and the increased utility.

3 Conclusions

3.1 Messages from the census data

The first message from the graphs above is about redesign of geographical classifications. Our example and others can suggest where the minimum size should be set, and they can indicate what the effect of this is on utility and safety. The upward pressure on this number needs to be balanced by downward pressure from user needs, and from geographical practicalities.

The second message is about classifications, and will become clearer with further examples. Some classifications are appropriate for certain tables, and some are not. The Age classification that we used (15–19, 20–24 to 65+) is good for the employed population. An alternative (0–9, 10–19 to 100+) would give tables where over half the cells held very few, if any, employed people. We need to design a variety of classifications for different output needs, and test them against the data with these graphical tools.

The third message is about safety and utility. For safety, our complex rule set, RMT, deals with sparseness to the left of the limit in Fig 2a, but not to the right. For utility, Fig 2b and all of Fig 3 show that RMT is worthwhile, both in releasing counts for high proportions of persons and in limiting the noise added to counts.

3.2 Further issues for investigation

For our 2011 Census, we are open to all possibilities for data access. We need to assess our RMT path against the quite different paths being developed and taken by other agencies. We need to deal with means of numerical variables, and other derivations, in consistent ways. We need to continue and enhance our microdata access: via licensed files, remote access, and our three data laboratories. Along with all this, we need a programme to inform users of these options, and enable them to strengthen their methods of accessing our data.

3.3 Summary

If we wish to optimise both utility and safety for tables of counts, we need to act at the design stage of the statistical process, as well as at the dissemination stage. The design of geographical and other classifications needs to combine evidence from within a census dataset with the needs of users. At the dissemination stage, it is appropriate to respond to a situation of complex risks and user needs with a complex set of rules that can filter out for release the information that is most useful and safe.

References

Duncan, G.T., Fienberg, S.E., Krishnan, R., Padman, R. & Roehrig, S.F. Disclosure Limitation Methods and Information Loss for Tabular Data. (2000). In *Confidentiality, Disclosure, Data Access: Theory and Practical Applications for Statistical Agencies.* Doyle, P., Lane, J.I., Theeuwes, J.M.M., & Zayatz, L.M. (Eds). Elsevier Science.

Jackson, L.F., Corscadden, L., & Zeng, I. (2005). *Small Area Disclosure Control: When Do Uniques Occur?* The 55th Session of the International Statistical Institute Sydney 2005 Proceedings.

Schlomo, N. (2005). *Assessment of Statistical Disclosure Methods for the 2001 UK Census*. Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality. Geneva.

Schlomo, N., & Young, C. (2005). *Information Loss Measures for Frequency Tables*. Joint UNECE/Eurostat work session on statistical data confidentiality. Geneva.

Wooton, J., & Fraser, B. (2005). *A Review of Confidentiality Protection for Statistical Tables with Special Reference to the Differencing Problem*. Australian Bureau of Statistics, Canberra.