Working Paper No. 28
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 ECE/Eurostat work session on statistical data confidentiality
(Luxembourg, 7-9 April 2003)

Topic (i): New theories and emerging methods

THE NOISE METHOD FOR TABLES - RESEARCH AND APPLICATIONS AT STATISTICS NEW ZEALAND

Contributed Paper

Submitted by Statistics New Zealand[1]

[1] Prepared by Mike Camden, Katrina Daish and Frances Krsinich. Acknowledgements to Mike Doherty and Jamas Enright.

Introduction

Tables of business magnitude data, which are the main output from business surveys, have a relatively high disclosure risk. Business populations tend to be skewed, with many small and medium-sized businesses and just a few very dominant businesses. This means that, for sampling efficiency, the large businesses are usually in full-coverage strata, so there is no confidentiality protection from sampling for these units.

There is generally good public knowledge about the industry, size and region of businesses, and their approximate market share. This information can enable close approximations of confidential information for those cells dominated by just a few large businesses. In particular businesses can use their own data to deduce the characteristics of other businesses in the same cell.

Cell suppression is a common method for protecting tables of business magnitude data. Cell suppression is a two-stage process. A dominance rule, such as the *(n,k)* rule, is used to identify sensitive cells for suppression. Suppression of these sensitive cells gives the 'primary suppressions'. To protect against derivation of the sensitive cells by subtraction from the marginal totals of the table, 'secondary suppressions' of non-sensitive cells are usually also required.

Determination of the secondary suppression patterns is non-trivial, particularly for large and complex tabulations. Upper and lower bounds for the suppressed cell values can be derived by solving the equations implied by the interior and marginal cell values, in addition to non-negative constraints on the cell value (Cox, 1995). The intervals within these bounds are referred to as ‘feasibility intervals’ for the suppressed value. An optimal secondary suppression pattern minimises the information loss due to cell suppression while ensuring that feasibility intervals around sensitive values are sufficiently large to preserve respondent confidentiality.

The Noise Method as an Alternative

In 2000, analysts of the Workplace Disputes Survey, a survey run jointly between Statistics New Zealand and the New Zealand Department of Labour, had difficulty applying random rounding and the *(n,k)* rule with their output software. They suggested adding unbiased random noise at the unit record level instead. After consideration Statistics New Zealand approved this as an adequate confidentiality measure, particularly since the sample design was non-standard (for a business survey), in that not all large businesses were full-coverage - it therefore had less risk associated with it than a standard business survey. Sampling weights were 'disturbed' by an amount inversely proportional to the sampling weights, to adjust for the protection already offered by sampling.

Later in 2000, Laura Zayatz from the US Census Bureau gave a paper titled "Using Noise for Disclosure Limitation of Establishment Tabular Data" at the Second International Conference on Establishment Surveys (ICES2) in Buffalo, New York (Zayatz et al 2000). This outlined a similar method. Rather than disturbing the weights directly, a 'multiplier' is generated. The multiplier is only applied to the sampled unit (and not to those other units in the population which the sampled unit represents), which results in the level of disturbance being inversely proportional to the weight. The method was experimentally applied to the US Census Bureau's

Research and Development Survey, and various summary statistics were computed to empirically test the properties of the method.

We replicated this work using our own Annual Enterprise Survey (AES) data, and the results are presented in a Statistics New Zealand research report (Krsinich and Piesse, 2002). Our results were very similar to Zayatz et al (2000). In addition, we defined and computed some information loss measures to compare cell suppression and the noise method for the tables we were working with.

How the Noise Method Works

For each observation, or unit, in the data, a multiplier is randomly generated from some distribution centered around 1. A bimodal distribution with all values at least, say, 10% away from 1 ensures that each unit's value is perturbed by at least 10%.

Before tabulation, values are multiplied by (multiplier + (weight - 1)), rather than by their original sampling weight.

The method is unbiased (Zayatz etal 2000; Evans, Zayatz and Slanta 1998). That is, the expected value of the 'noised' cell is equal to the original cell value.

The method is illustrated below in tables 1 to 4 with a simple example:

Table 1. Fictional microdata and multipliers

id	Industry	Region	Turnover ($000)	Weight	Weighted value	Multiplier	'Noised' weighted value
1	A	a	50	1	50	1.12	56 $(=50\times1.12)$
2	A	b	30	1	30	1.09	32.7
3	A	b	40	1	40	1.11	44.4
4	B	a	12	5	60	0.91	58.92 $(=12\times4.91)$
5	B	a	14	5	70	1.10	71.4
6	B	b	7	100	700	0.88	699.16 $(=7\times99.88)$
7	B	b	2	100	200	0.93	199.86
8	B	b	3	100	300	1.11	300.33
9	B	b	4	100	400	0.90	399.6

Table 2. Original table - Turnover ($000)

	Region a	Region b	Total
Industry A	50	70	120
Industry B	130	1600	1730
Total	180	1670	1850

Table 3. Noised table - Turnover ($000)

	Region a	Region b	Total
Industry A	56	77.1	133.1
Industry B	130.32	1598.95	1729.27

Total	186.32	1676.05	1862.37

Table 4. Percentage difference between noised and original table

	Region a	Region b	Total
Industry A	12	10	11.0
Industry B	0.2	-0.07	-0.04
Total	3.5	0.3	0.6

The percentage difference (i.e. the 'noise') is greater for full-coverage cells. In our example, cells corresponding to *Industry A* both consist solely of full-coverage businesses (i.e. weights = 1), with one business in *Region a* and two in *Region b*. These have 12% and 10% noise respectively. On the other hand, *Industry B, Region a* has 2 medium-sized businesses (weights = 5) and receives 0.2% noise. *Industry B, region b* has 4 small businesses (weights = 100) and receives only 0.07% noise.

It is important to note that, although it is *applied* at the microdata level, the noise method *protects* the table, not the microdata.

The Advantages of the Noise Method

For any dataset, the noise method is applied only once. From then on, all tables produced from the dataset will be consistent, both internally (i.e. tables will be additive) and externally (i.e. related tables will be consistent with each other). A consequence of this is that there are no disclosure risks posed from multiple production of the same table, or production of related tables. Also, the size and/or complexity of tables doesn't affect the application of the method.

Publication of noised sensitive cells gives the users an approximate value, rather than complete suppression.

In general, more noise is added to the sensitive cells, less noise is added to the non-sensitive cells. So the information loss is being targeted to those cells which pose the most risk.

Simulation Study Following the Approach of Zayatz et al (2000)

The Annual Enterprise Survey (AES) is Statistics New Zealand's largest financial survey. It was recently redesigned (Krsinich 2000) and now makes extensive use of tax data for small and simple businesses. The post out sample is around 20,000, with full coverage of approximately another 200,000 units whose data are derived directly from tax data.

Despite the large sample sizes resulting from full coverage of small businesses via the use of tax data, sensitive cells still occur in the standard published AES tables. Together with the secondary suppressions, there can be a significant loss of information resulting from cell suppression in some of the standard published AES tables.

We used AES 1999 data as an example for trialling the noise method and followed the same approach as Zayatz et al (2000) to ensure direct comparability with their work. In addition to this, we discussed and defined some measures of information loss to enable a comparison of information loss between cell suppression as currently performed in AES, and the noise

method. Krsinich and Piesse (2002) gives a more detailed account of this work than we have space for here.

Three industries with suppressions in the standard published tables for the variable Total Income were chosen. Note that these are examples of AES tables with relatively high levels of suppressions. Primary suppressions are indicated by a 'p' and secondary suppressions by an 's' in tables 5 to 7 below.

Each of the tables forms part of what is effectively a 4-dimensional table, with relationships across time, and all-industry totals, as well as the more obvious 1-digit industry totals and total income marginals corresponding to the two-dimensional tables below.

Table 5. Industry F (Wholesale): 3-digit level

Industry	sales - nfp	sales - other	interest	govt fund	non-op	total inc
F011						
F012						
F013				s	s	
F014						
F015				p	s	
F016				s	s	
F017						
Total F						

Table 6. Industry K (Finance and Insurance): 2-digit level

Industry	sales	interest	govt fund	non-op	total inc
K01			s	s	
K02			p	s	
K03					
Total K					

Table 7. Industry P (Cultural and Recreational Services): 3-digit level

Industry	sales	interest	govt fund	non-op	total inc
P011		p	s		
P012		s	s		
P013		s	s		
Total P					

We applied approximately 10% noise to each unit's value.

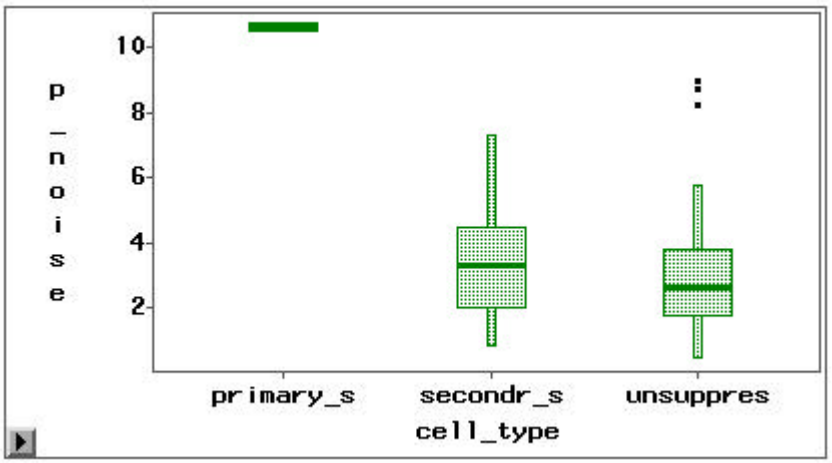
There is some hierarchical structure to the sampled units. More than one unit can belong to the same 'group'. To ensure protection at the group level, two stages of randomisation are used. The first assigns a 'direction' at the group level - that is, each group of units has a 0.5 probability of having a multiplier close to 0.9 and a 0.5 probability of having a multiplier close to 1.1. In the second stage, the units within each group are assigned a multiplier from the distribution around whichever of 1.1 or 0.9 has been assigned to its group.

Then, for every unit, the values of interest are multiplied by (multiplier+(weight-1)) before tabulation.

We ran 1000 replications of the three noised AES tables, and computed summary statistics to describe the behavior of the cells across the replications.

We computed the absolute percentage noise in each cell for each replication. We then averaged these absolute percentages across the 1000 replications. The distribution of this 'average absolute percentage noise' across cells of different types is shown below in graph 1.

Graph 1. Average absolute percentage noise



Primary suppressions (3 cells)
Av 10.7, Med 10.6, Max 10.8, Min 10.5

Secondary suppressions (13 cells)
Av 3.5, Med 3.3, Max 7.3, Min 0.87

Unsuppressed cells (72 cells)
Av 3.0, Med 2.6, Max 9.0, Min 0.46

Primary suppressions are those cells which are defined as sensitive by Statistics New Zealand's version of the *(n,k)* rule. As expected, these receive significantly more noise than the non-sensitive cells. We want these cells to receive more noise, because these are the cells we want to protect. Conversely the presence of significantly less noise in the non-sensitive cells (i.e. both the 'secondary suppressions' and the 'unsuppressed cells' in Graph 1) is a desirable result as these are the cells we don't need to protect against disclosure.

Information Loss Comparisons

The noise method results in the addition of at least some noise to every cell. The cell suppression method results in complete suppression of primary and secondary suppressed cells, but other cells are left unchanged. We discuss, define and compute some measures of information loss due to cell suppression, and we then compare these to the average absolute percentage of noise due to the noise method.

When considering the protection offered by a method such as cell suppression, we assume that an intruder can, and will, combine the equations implied by the remaining values in the tables, to derive feasibility intervals for each suppressed value.

While this worst case scenario might be a necessary assumption for guaranteeing a specified level of disclosure limitation, it is perhaps more useful to consider a 'non-intruding user' when trying to quantify information loss. We therefore compute two different information loss measures, corresponding to both the 'intruder' and 'non-intruder' scenarios.

For the ‘intruder scenario’ we calculate the information loss corresponding to the feasibility intervals resulting from the particular cell-suppression pattern that was used for AES99.

But most users won't be able and/or willing to put in the work necessary to derive feasibility intervals, particularly for large or complex linked or hierarchical tables. For these users, we assume that the information lost due to cell-suppression is the full value of the cell.

The Intruder's Information Loss from Cell Suppression

We calculated the intruder’s information loss as the half-width of the feasibility interval divided by the midpoint of the feasibility interval. See Krsinich and Piesse (2002) for a discussion of why we adopted this particular measure. We calculated the 'intruder's information loss' for the three AES99 tables being considered and, from these, we calculated the average intruder's information loss for each type of cell, to compare to the average absolute percentage noise using the noise method. The results are shown in table 8.

Table 8. Information loss comparison – Intruder scenario (%)

Type of cell	Cell suppression	Noise method
Primary suppression (3 cells)	100	11
Secondary suppression (13)	61	3.5
Interior cell (59)	19	3.7
Unsuppressed cells (72)	0	3.0
All cells (88)	12	3.3

The User's Information Loss from Cell Suppression

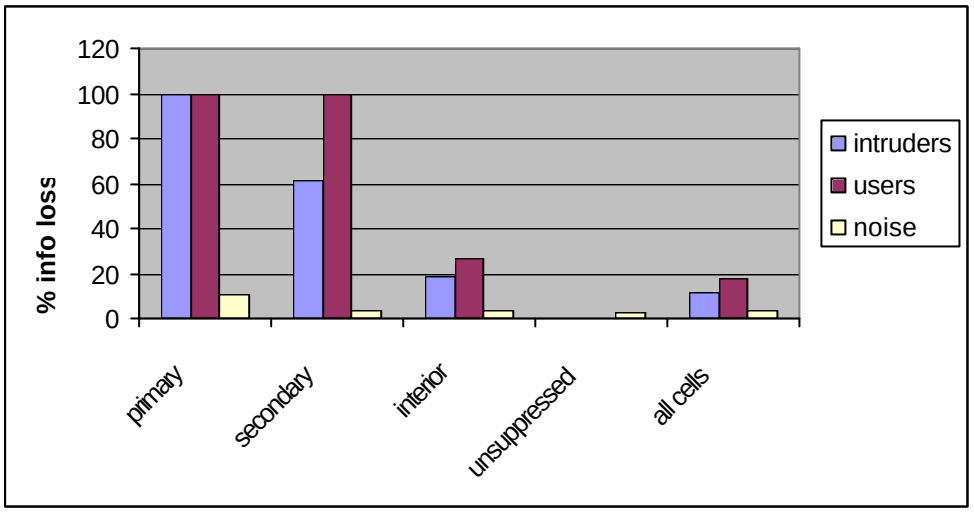
As discussed above, it could be argued that, for most users, a suppressed cell represents a complete loss of information for that cell. So, using a simpler information loss measure which measures each suppressed cell as a 100% loss of information, we can derive the average 'user's information loss' for each different type of cell.

Table 9. Information loss comparison – User scenario (%)

Type of cell	Cell suppression	Noise method
Primary suppression (3 cells)	100	11
Secondary suppression (13)	100	3.5
Interior cell (59)	27	3.7
Unsuppressed cells (72)	0	3.0
All cells (88)	18	3.3

Using either information loss measure, it can be seen that the noise method compares very favourably in terms of the average amount of information lost for these tables. The trade-off is that cells which would have been unchanged under the cell-suppression method, are now ‘noised’. Graph 2 shows the comparison visually.

Graph 2. Information loss comparison



Extensions to the Method for Implementation for AES 2003

Given the promising results from the research outlined above, we hope to implement the noise method for our Annual Enterprise Survey in 2003.

Further testing for this planned implementation has used AES 2000 data, and a multiplier derived from a truncated half-normal distribution 10% away from 1. With a standard deviation of 0.02, the distribution is truncated at 0.1. This means that no business, or unit, in the survey has its response 'noised' by greater than 20% or less than 10%, and 95% of the businesses have somewhere between 10% and 15% noise added to, or subtracted from, their original value. We ran tests on the AES 2000 data using this distribution and the results reflected those of the earlier simulation study, which used a Beta distribution as in Zayatz et al (2000).

The main issue that has arisen for implementation is the potential effect of the noise method on estimates of movements, which are important survey outputs. The noise method would cause too much volatility in movements if it was applied independently each year. Therefore we have split the noise into two parts. The 'base noise' is the direction of the noise – that is, whether we add or subtract approximately 10%. A unit will retain its base noise for its life in the survey. We also apply 'extra noise' (from the corresponding truncated half-normal distribution) independently from year to year.

We have formulated the relative ‘sampling plus noise’ error for estimates of levels (see the Appendix). This remains to be done for movements. This information will be important for the ultimate decision on whether to adopt the method for AES 2003.

For a range of industries, and for three different survey variables (Total Income, Total Expenditure and Total Assets) we calculated the relative sampling-plus-noise error for the AES 2000 data, and compared this to the relative sampling-only error. This is shown in Table 10 below.

- Horticulture and Fruit Growing (HFG) has 12,671 units in the AES 2000 sample, and has no cell suppression.
- Other Food Manufacturing (OFM) has 352 units, and no cell suppression.
- Basic Metals (BM) has 131 units in the sample, and has some confidential cells.

- Telecommunication Services (TS) has 67 units, and is an industry that is dominated by a single large unit, therefore has many confidential cells.

Table 10. Relative sampling error and relative sampling-plus-noise error (%)

	Industry:	HFG	OFM	BM	TS
Total Income					
	Sampling	0.0	5.3	3.8	1.4
	Sampling plus noise	0.8	6.5	11.0	9.4
Total Expenditure					
	Sampling	0.0	5.3	3.5	1.8
	Sampling plus noise	0.4	6.5	11.3	17.1
Total Assets					
	Sampling	0.0	26.2	13.6	3.1
	Sampling plus noise	0.4	26.8	18.4	19.6

Extending the Noise Method to Tables of Counts

Count data can be considered to be magnitude data where every respondent, or unit, contributes a magnitude of one. Tables of counts from household survey data are generally protected by uniformly small sampling fractions, so we are only interested in the case of tables of counts from a census, where the implicit weight is one for every unit. By randomly perturbing the weight of 1 either up (to 2) or down (to 0), we can introduce confidentiality protection while preserving the statistical properties of the data.

This is a very simple and elegant approach which could be useful in situations where random rounding is not appropriate, such as when it is possible for an intruder to obtain many repetitions of the same, independently rounded, counts. This could arise in a remote access situation, or if a full suite of many-dimensional tables with shared, independently rounded, marginals was produced. That is, the noise method for counts would avoid problems of undoing protection via comparison of related tables.

A potential problem with the method is the level of variance introduced for large-valued cells. If the probability of perturbing the weight from 1 to 0 is p, and the probability of perturbing the weight from 1 to 2 is also p (i.e. the method is unbiased), then the variance of the introduced noise is $2pc$, where c is the unperturbed cell count. So, for example, with a p of 0.2, and a cell value of 1000, the variance introduced is 400 which translates to a 'relative noise error' of $1.96\sqrt{400}/1000 = 3.9\%$.

This has led us to consider various alternatives based on using the 'noised' cell under a certain cell size threshold, and the original cell over that threshold. These alternatives are not as simple or elegant, and suffer from non-additivity and some small potential for 'unlocking' - similar to the small number of unusual examples that can be 'unlocked' under random rounding. However, depending on the average size of the cells, whether there is potential for many related tables to be produced, and the relative strengths and weaknesses of the alternative methods available, the noise method for counts may be a useful approach to consider.

Conclusion

The noise method is very promising for tables of magnitudes from business surveys. We will soon decide whether the method will be officially used for the 2003 AES survey results. Our experiences in operationalising the method may prove useful to other agencies considering adopting the noise method in the future.

There appears to be good potential for extending the method to tables of counts, and we hope this will be explored further, either by ourselves or others.

References

Cox, L H (1995) 'Protecting confidentiality in business surveys' in *Business Survey Methods*, B G Cox et al (eds), 443-476, Wiley, New York.

Evans T, Zayatz L and Slanta T (1998), *Using Noise for Disclosure Limitation of Establishment Surveys*, Journal of Official Statistics, Vol 4, No. 4.

Krsinich F (2000), *Tax Data in Statistics New Zealand's Main Economic Survey: A Two-Phased Redesign*, Proceedings of the Second International Conference on Establishment Surveys, Buffalo, New York.

Krsinich F and Piesse A (2002), Multiplicative Microdata Noise for Confidentialising Tables of Business Data, Research Report # 19, Statistics New Zealand. Available online at www.stats.govt.nz via 'publications' then 'technical publications'.

Zayatz L, Evans T and Slanta J (2000), *Using Noise for Disclosure Limitation of Establishment Tabular Data*, Proceedings of the Second International Conference on Establishment Surveys, Buffalo, New York.

Appendix. Relative Sampling-Plus-Noise Error - Formulation and Example

$$v_{s,n} = v_s + n_n \sum_{i=1}^{N} p_i y_i^2 \quad \text{.....(1)}$$

Where $v_{s,n}$ is the variance due to both sampling and noise, v_s is the variance due to sampling and v_n is the variance due to noise.

For the truncated half normal distribution described in this paper, $v_n = 0.0146$.

p_i is the probability that business i was selected into the sample. Note that $p_i = \frac{1}{w_i}$, where w_i is the weight for business i.

y_i is whatever is being estimated (e.g. 'Total Income') for business i

The sum is across the population (i.e. i = 1 to N)

We only have sample data, so the formula needs to be restated in terms of the sample.

Note that the sum across the population can be estimated by the weighted sum across the sample. That is, $\sum_{i=1}^{n} w_i p_i y_i^2 \approx \sum_{i=1}^{N} p_i y_i^2$ and $\sum_{i=1}^{n} w_i p_i y_i^2 = \sum_{i=1}^{n} w_i \frac{1}{w_i} y_i^2 = \sum_{i=1}^{n} y_i^2$ (from above), so we can restate formula (1) as $v_{s,n} = v_s + v_n \sum_{i=1}^{n} y_i^2$ .....(2)

From the relative sampling errors (RSE) produced for the AES estimates we can derive $v_s = \left(\frac{RSE \times estimate}{1.96} \right)^2$

For example, we have an estimate of 5.3% for the relative sampling error of the Total Income estimate for the industry Other Food Manufacturing (OFM).

The Total Income estimate for industry OFM is 5,510,151 ($000)

so, for this example, $\text{var}_s = \left(\frac{0.053 \times 5{,}510{,}151}{1.96} \right)^2 = 2.22 \times 10^{10}$

From the sample data we can calculate $\sum_{i=1}^{n} y_i^2 = 79.44 \times 10^{10}$

and we already have that $v_n = 0.0146$

So, substituting this into (2), we get $v_{s,n} = 2.22 \times 10^{10} + 0.0146 \times 79.44 \times 10^{10} = 3.38 \times 10^{10}$

Stating this in terms of relative sampling-plus-noise error, for comparison with the sampling error:

$$RSNE = \frac{1.96\sqrt{3.38 \times 10^{10}}}{5{,}510{,}151} = 0.065 = 6.5\%$$

So, for the variable Total Income for industry OFM (Other Food Manufacturing) in AES 2000, our use of the noise method means that the relative sampling-plus-noise error is **6.5%**, compared to the relative sampling error of **5.3%**.