

Multiplicative Microdata Noise for Confidentialising Tables of Business Data

**Application to AES99 data
with a comparison to cell suppression**

**Frances Krsinich and Andrea Piesse
Analytical Support**

**Statistics New Zealand
January 2002**

Published in January 2002 by

Statistics New Zealand
Te Tari Tatau
Wellington, New Zealand

Catalogue Number 01.095.000
ISBN 0-478-26900-5

Preface

Statistics New Zealand, like most other official statistical agencies, is legally and ethically obliged to protect against the disclosure of individuals' responses. Confidentiality rules, also known as statistical disclosure control methods, are necessary to ensure that individual responses can't be derived from tables of data. The cells which pose the greatest confidentiality risk are those corresponding to unusual or dominant respondents.

Multiplicative Microdata Noise for Confidentialising Tables of Business Data discusses a new method for confidentialising tables of business data called the 'noise method'. This method is slowly being adopted overseas and might be used for business data within Statistics New Zealand in the future. This report was written by Frances Krsinich and co-authored by Andrea Piesse under the direction of Sharleen Forbes, Chief Analyst.

I would like to acknowledge my appreciation to the authors, and to Senior Mathematical Statisticians Richard Penny and Richard Arnold, and Laura Zayatz from the US Census Bureau who reviewed this report and provided useful feedback, some of which has been incorporated, and some of which will inform future research. Also, I would like to thank the Methodological Advisory Committee of the Australian Bureau of Statistics, in particular David Steel, for providing useful feedback when the report was presented in November 2001.



Brian Pink
Government Statistician
STATISTICS NEW ZEALAND

Contents

	Page
1. Summary	7
2. Introduction	8
3. Cell Suppression in the Annual Enterprise Survey	9
4. The Noise Method	10
5. Application of the Method to Annual Enterprise Survey Data	14
6. Results of Analysis	15
7. Information Loss Comparisons	20
8. Future Work	23
9. Conclusions	25
References	26
Appendix - Computing Feasibility Intervals	27

Contents - continued

Tables

1	Fictional microdata and multipliers	12
2	Original table – turnover	12
3	Noised table – turnover	12
4	Percentage difference between noised and original table	12
5	Industry F (wholesale) – three-digit level	14
6	Industry K (finance and insurance) – two-digit level	14
7	Industry P (cultural and recreational services) – three-digit level	14
8	Industry F (wholesale) – conditional coefficients of variation	15
9	Industry K (finance and insurance) – conditional coefficients of variation	16
10	Industry P (cultural and recreational services) – conditional coefficients of variation	16
11	Counts of cells according to amounts of noise – large (> 4%) or small (< 4%)	19
12	Average intruder’s information loss – comparison of cell suppression to the noise method	21
13	Average user’s information loss – comparison of cell suppression to the noise method	21
A1	Fictional unconfidentialised table	27
A2	Fictional table with cell suppressions applied	27

Charts

1	Beta distributions used to generate multipliers at least 0.1 away from 1	11
2	Boxplots of the conditional coefficient of variation for different types of cells	16
3	Average absolute percentage noise – all cells	17
4	Average absolute percentage noise – interior versus marginal cells	17
5	Average absolute percentage noise – primary and secondary suppressed cells versus unsuppressed cells	18

Figures

1	Comparison of information loss between cell suppression (for intruders and users) and noise method	22
---	--	----

1. Summary

This paper discusses problems with the current cell suppression method of confidentialising business survey data, introduces a new method called the 'noise method' and presents the results of applying this new method to the Annual Enterprise Survey (AES).

We follow the approach taken by the US Census Bureau, who applied the 'noise method' to their Research and Development Survey. We get a very similar pattern of results, although the average level of noise is lower.

We then develop a measure of information loss to compare the noise method with cell suppression, and get very positive results.

Finally, we discuss areas for further research and consideration and, in particular, we outline how the method can be extended to tables of counts, such as those produced from the Population Census.

2. Introduction

Statistics New Zealand, like most other official statistical agencies, is legally and ethically obliged to protect against the disclosure of individuals' responses. An 'individual' could be a person, a business or a household. Confidentiality rules, also known as statistical disclosure control methods, are necessary to ensure that individual responses can't be derived from tables of data. The cells which pose the most confidentiality risk are those corresponding to unusual or dominant individuals.

Tables can be categorised into tables of counts (for example, number of businesses in each industry and region) and tables of magnitudes (such as the total turnover of businesses in each industry and region).

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. This contrasts with most household sample surveys where, without detailed information about the sample design, sample membership cannot be established with certainty. Also, 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 one or two 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. It is a two-stage process. A dominance rule, such as the (n,k) rule, is used to identify sensitive cells for suppression. These cells are called 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 required. Determination of the secondary suppression patterns is complicated, 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. The intervals within these bounds are referred to as 'feasibility intervals' (see the Appendix for an example of how these are calculated). An optimal secondary suppression pattern would need to minimise the information loss due to cell suppression while ensuring that feasibility intervals around sensitive cells are sufficiently large.

3. Cell Suppression in the Annual Enterprise Survey

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 is derived directly from tax data.

Statistics New Zealand's Business Frame uses the following hierarchical structure for its statistical units: Group Top Enterprises (GTEs) consist of one or more Enterprises (ENTs), which consist of one or more Kind-of-Activity units (KAUs), which consist of one or more Geographic Units (GEOs).

The statistical unit in AES is the Kind-of-Activity unit. Note that the application of the noise method utilises the GTE identifiers to ensure that each group of KAUs within a GTE receives noise in the same 'direction' (in other words, the random noise is either added or subtracted from all KAUs within a GTE to ensure that approximately 10% noise results at the GTE level).

Despite the large sample sizes resulting from full coverage of small businesses, 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 these tables.

Statistics New Zealand's Business Statistics division recently did some work to determine ways of reducing the impact of cell-suppression on the published results from AES (Pang 2001). Approaches considered included the reformatting of standard tabulations, relaxing the level of disclosure control on older data; changing from the (n,k) rule to the $p\%$ rule (a less conservative dominance rule than the (n,k) rule, which incorporates information about the distribution of the cells); using automated cell suppression software; changing the procedures for determining secondary suppressions; and putting more resource into determining what data is already publicly available.

3.1 The (n,k) rule

Statistics New Zealand currently uses the (n,k) rule to determine primary suppressions in business magnitude data. The (n,k) rule is one of a class of linear sensitivity measures for determining sensitive cells. A cell is defined as sensitive if n or fewer respondents contribute at least $k\%$ to the cell total.

The (n,k) rule ensures that any cell in which a respondent's value can be estimated by a 'coalition' of size $n-1$ to within $(100-k)/k\%$ will be defined as sensitive. So, for example, a $(3,70)$ rule would ensure that no cell could be estimated to closer than 43% by a coalition of size 2. The (n,k) rule is simple to apply, as it only requires knowledge of the total cell value and the values of the n largest contributors to that cell. The trade-off for this simplicity is that it tends to oversuppress because it doesn't take into account the distribution of values within the cell and therefore must assume a worst-case scenario. For example, the $(3,70)$ rule guarantees that any cell which can be estimated to closer than 43% by a coalition of 2 will be defined as sensitive, but it doesn't guarantee that a cell which cannot be estimated that closely won't also be defined as sensitive.

The exact parameters of the (n,k) rule are kept confidential to avoid compromising the effectiveness of the rule (Willenborg and de Waal (1996), 107). Usually, n is 2, 3 or 4 and k is > 70 .

3.2 Secondary suppression

The main difficulty with the current cell-suppression method is secondary suppression. Note that secondary suppression is also sometimes referred to as 'consequential suppression', or 'complementary suppression'. Secondary suppression is difficult to apply, and it results in the loss of non-sensitive cells. An optimal secondary suppression pattern is one in which feasibility intervals are sufficiently wide, and information loss is minimised. We use 'information' to refer to the usefulness of the data to answer questions. The definition of information loss is problematic. Ideally, an efficient measure of information loss depends on the intended use of the data, but in practice this ideal is infeasible due to the varying uses made of Statistics New Zealand data. Usually, in automated cell-suppression packages, information loss is defined to be proportional to the number of cells suppressed, or to the total value of cells suppressed. In general, the secondary cell suppression problem is difficult to automate – solution procedure complexities grow exponentially in relation to the size of the problem or table. Also, it is necessary to keep track of which cells have been suppressed in previously released related tables. Because of the computational intensity, it is common for heuristic methods to be used in automated packages, rather than provably optimal procedures. Manual secondary suppression (which is what Statistics New Zealand does currently) can be error prone and very time consuming, particularly for large and/or complex tables.

4. The Noise Method

4.1 Background

In 2000, the market research company ACNielsen was the analyst for the Workplace Disputes Survey, a survey run jointly between Statistics New Zealand and the Department of Labour. Their output software was such that random rounding and the (n,k) rule were going to be difficult to apply, and so they suggested adding unbiased random noise at the unit record level instead. After some 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 and therefore the survey had less risk associated with it than a standard business survey. Sampling weights were 'disturbed' by an amount inversely proportional to themselves, to adjust for the protection already offered by sampling. A 'disturbance' changes a cell by a certain amount up or down – in a sense it is adding variation. At that stage, Statistics New Zealand was not in a position to investigate the analytical implications of the technique but saw that it was a very promising idea, well worth pursuing further for our own use.

Later in 2000, Laura Zayatz from the US Census Bureau gave a paper "Using Noise for Disclosure Limitation of Establishment Tabular Data" at the second International Conference on Establishment Surveys (ICES2) (Zayatz, Evans and Slanta 2000). This outlined a method essentially equivalent to the one suggested by ACNielsen. However, 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 decided to replicate this work using our own Annual Enterprise Survey (AES) data, and the results are presented in this paper.

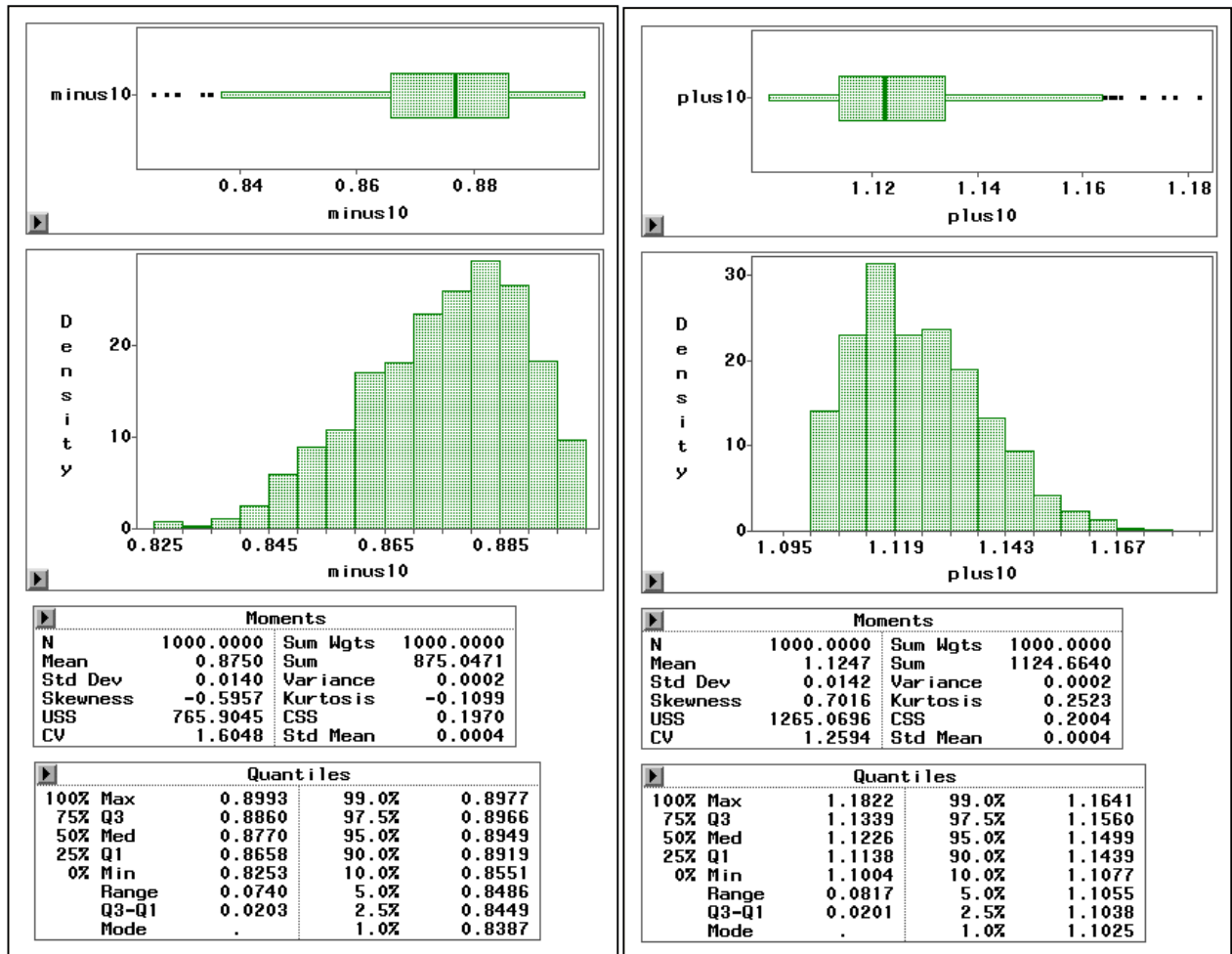
Not long ago Julia Lane from the US Census Bureau came to Statistics New Zealand during her visiting research fellowship at the Department of Labour. She told us that the noise method has recently been approved by the US Census Bureau's Disclosure Review Board for use in the Longitudinal Employer-Household Dynamics Program – a major integration of individual and business data.

4.2 How the method works

For each observation in the data, a multiplier is randomly generated from some distribution centered around 1. For our work (following Zayatz, Evans and Slanta 2000), the multiplier is Beta-distributed bimodally about 0.9 and 1.1, using the distributions $0.1 * \text{Beta}(6,2) + 0.8$ and $0.1 * \text{Beta}(2,6) + 1.1$ respectively. Before tabulation, values are multiplied by a function of this multiplier. Chart 1 shows the Beta distributions used.

Chart 1

Beta Distributions Used to Generate Multipliers at Least 0.1 Away From 1



Values are multiplied by the randomly generated multiplier added to the sampling weight less one, rather than by just their original sampling weight:

Data value without noise = Original data value * sampling weight
 Data value with noise = Original data value * (multiplier + (sampling weight – 1))

So large businesses with weights of 1 will be multiplied by approximately either 0.9 or 1.1, while small businesses with weights of, say, 100 will be multiplied by approximately either 99.9 or 100.1, instead of by their original weights of 100. The use of the Beta distribution ensures that the multiplier is at least 10% different from 1. Note that the use of the term ‘multiplier’ is a bit confusing, as the unweighted value is multiplied by a function of the multiplier, rather than the multiplier itself. We use the term for consistency with the US Census Bureau’s work.

The method is unbiased (Zayatz, Evans and Slanta 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 with a simple example, shown below in Tables 1 to 4:

Table 1

Fictional Microdata and Multipliers

Obs	Industry	Region	Turnover	Weight	Weighted Value	Multiplier	Noised Weighted Value
			\$(000)				
1	A	a	50	1	50	1.12	56 (=50*1.12)
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*4.91)
5	B	a	14	5	70	1.1	71.4
6	B	b	7	100	700	0.88	699.16 (=7*99.88)
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	0.9	399.6

Table 2

Original Table Turnover

	Region a	Region b	Total
	\$(000)		
Industry A	50	70	120
Industry B	130	1600	1730
Total	180	1670	1850

Table 3

Noised Table Turnover

	Region a	Region b	Total
	\$(000)		
Industry A	56	77.1	132.3
Industry B	130.32	1598.95	1729.27
Total	186.32	1675.25	1861.57

Table 4

Percentage Difference Between Noised and Original Table

	Region a	Region b	Total
	Percent		
Industry A	12	10	10.3
Industry B	0.2	-0.06	-0.04
Total	3.5	0.3	0.63

Note that the percentage difference (that is, the 'noise') is greater for full-coverage cells. Cells corresponding to *Industry A* both consist solely of full-coverage businesses (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 two medium-sized businesses (weights = 5) and receives 0.2% noise. *Industry B region b* has four small businesses (weights = 100) and receives only 0.06% noise.

4.3 The advantages of the noise method

Compared to the cell suppression method, the noise method is easy to understand, and simple to apply.

The size and/or complexity of tables doesn't affect the application of the method. Given the increasingly detailed data required of, and produced by, statistical agencies, this is an important feature.

For any dataset, the method is performed only once. From then on, all tables produced from the dataset will be consistent, both internally (tables will be additive) and externally (related tables will be consistent). A consequence of this is that there are no disclosure risks posed from multiple production of the same table (as is a possibility with remote access), or production of related tables.

In general, more noise is added to the sensitive cells, less noise is added to the non-sensitive cells. So the information loss occurs in those cells which pose a risk, with minimal disruption of the rest of the table. This frees the agency from having to make difficult, and often unavoidably arbitrary, decisions about which non-sensitive cells are going to be more or less useful for future research.

If the 'noised' sensitive cells are published, this serves as publication of an approximate value, rather than complete suppression. For the user, this approximate value is more useful than a complete suppression.

4.4 Classification of the noise method as a confidentiality method

An important point to note is that the noise method described above, although applied at the microdata level, is a method for protecting tables, not microdata. Statistics New Zealand is not currently able to produce microdata for public use, due to restrictions imposed by our Statistics Act, so this is an area of less immediate concern to us.

We searched for references to this general approach in the confidentiality literature and, in addition to the recent work by the US Census Bureau, we found two main references.

1. **Schlackis** (1993) discusses random perturbation, and distinguishes between output perturbation (adding noise to the cells of a table already produced from raw data), and input perturbation (adding noise to the raw data from which the tables are created). According to this classification, the noise method is a form of input perturbation, while random rounding to base 3 (used by Statistics New Zealand for tables of counts from the Population Census) would be classified as a form of output perturbation. Schlackis notes that there are problems with additive random perturbation which might be solved by using multiplicative random perturbation (as we do in our work), but doesn't pursue this any further.
2. **Willenborg and de Waal** (2000) briefly discuss 'source data perturbation' (SDP) as a method for producing tables which are free from the risk of a confidentiality breach. They note that either continuous variables could be perturbed, or the weights associated with each respondent (which equal 1, as they are considering the case of census data) could be perturbed. Willenborg and de Waal point out that the advantage of SDP is that consistency between cell values in different tables is guaranteed but, as in Schlackis (1993), the potential of SDP as a viable alternative to cell suppression isn't pursued any further in this work.

5. Application of the Method to Annual Enterprise Survey Data

5.1 US research

By taking the same approach as Zayatz, Evans and Slanta (2000), we hoped to ensure direct comparability with their work, with differences in results able to be attributed to the different data. In general, we would expect that the level of noise resulting would be inversely proportional to the homogeneity of the population (with respect to the variables being tabulated) and the number in the sample.

In addition to this, we discuss and define some measures of information loss to enable a comparison of information loss between cell suppression (as currently performed in AES) and the noise method.

5.2 Annual Enterprise Survey data used

We used AES99 data and chose three industries with suppressions in the standard published tables for total income. Primary suppressions are shown by a 'p' (these are the sensitive cells as defined by Statistics New Zealand's version of the (n,k) rule), and secondary suppressions are indicated by an 's'.

The relatively high number of secondary suppressions reflects the fact that each of the tables forms part of what is effectively a four-dimensional table, with relationships across time, and all-industry totals, as well as the more obvious one-digit industry totals and total income marginals shown here.

Table 5

Industry F (Wholesale)

Three-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)

Two-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)

Three-digit Level

Industry	Sales	Interest	Govt Fund	Non-op	Total inc
P011		p	s		
P012		s	s		
P013		s	s		
Total P					

6. Results of Analysis

Using the Beta distribution, as described above, we applied approximately 10% noise to each unit's value.

Two stages of randomisation were used. The first assigned a 'direction' at the group top enterprise (GTE) level – that is, each GTE had 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 Kind-of-Activity units (KAUs) within each GTE were assigned a multiplier from the Beta distribution around whichever of 1.1 or 0.9 had been assigned to the corresponding GTE.

For every KAU (that is, for every individual in the dataset), the values of interest were multiplied by (multiplier+(weight-1)).

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

6.1 The ratio of the average noised values to the original value

For each cell in the three tables, we calculated the ratio of a) the average of the 1000 noise-added values of the cell to b) the noise-free estimate. If the method is unbiased, we would expect this ratio to be very close to 1, for both sensitive and non-sensitive cells.

We calculated this ratio for all cells in the tables (including marginal totals) and all the values were very close to 1. The most extreme values (those furthest from 1) were **1.002** and **0.997**. This appears to empirically confirm that, as we expected, the noise method is unbiased.

6.2 The conditional coefficients of variation

We then looked at the standard deviation of the 1000 replications of the noised values for each cell. These were standardised by dividing by the true cell value. If the original cell value is considered to be fixed in terms of the addition of noise then the standard deviation of the noised cell values is the standard deviation of the noise itself. We use Zayatz et al's (2000) term – the 'conditional coefficients of variation' – to refer to these measures.

Cells currently suppressed are in bold and bordered, with the primary suppressed cells shaded and the secondary suppressed cells unshaded.

Table 8

Industry F (Wholesale)

Conditional Coefficients of Variation

Industry	Sales np	Sales Other	Interest	Govt Fund	Non-op	Total Inc
F011	0.02	0.04	0.08	0.04	0.03	0.02
F012	0.02	0.05	0.05	0.09	0.06	0.03
F013	0.02	0.02	0.03	0.01	0.03	0.02
F014	0.01	0.02	0.06	0.00	0.02	0.01
F015	0.02	0.04	0.06	0.11	0.05	0.02
F016	0.05	0.03	0.06	0.01	0.05	0.04
F017	0.01	0.01	0.01	0.09	0.03	0.01
Total F	0.02	0.02	0.03	0.05	0.02	0.02

Table 9

Industry K (Finance and Insurance)

Conditional Coefficients of Variation

Industry	Sales	Interest	Govt Fund	Non-op	Total Inc
K01	0.04	0.03	0.04	0.02	0.03
K02	0.03	0.05	0.11	0.08	0.04
K03	0.01	0.04	0.06	0.03	0.01
Total K	0.02	0.03	0.04	0.03	0.03

Table 10

Industry P (Cultural and Recreational Services)

Conditional Coefficients of Variation

Industry	Sales	Interest	Govt Fund	Non-op	Total Inc
P011	0.03	0.11	0.07	0.06	0.04
P012	0.01	0.02	0.07	0.04	0.03
P013	0.06	0.01	0.04	0.06	0.05
Total P	0.03	0.04	0.05	0.04	0.03

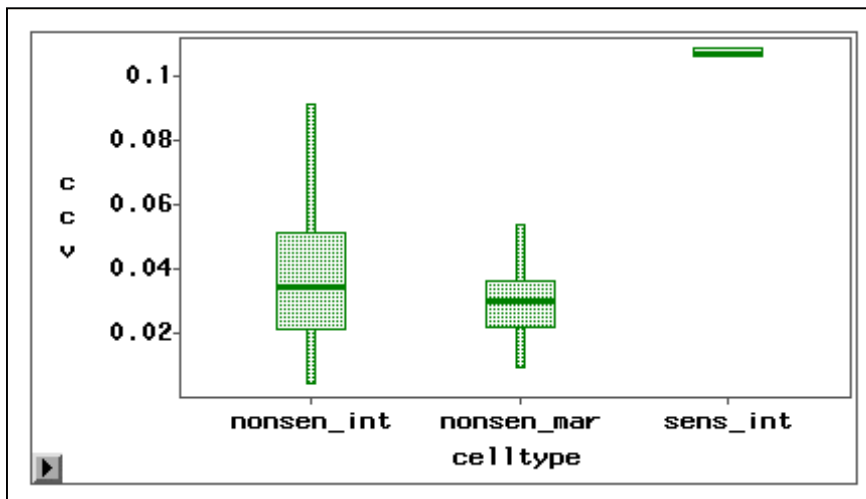
Note that the sensitive cells (the shaded primary suppressed cells) have significantly higher coefficients of variation than the other cells, including the secondary suppressions. This is as expected. The sensitive cells are those which either have very few respondents, or a few very dominant respondents, or both. If there are only a few respondents then their values (and the corresponding noise components of these values) are high relative to the cell total. Also, with only a few respondents there is less chance for the noise to 'balance itself out'. Similarly, if there are many respondents in the cells, but just a few of them are very dominant, then the values and noise of the dominant respondents are relatively large compared to the cell total. There is less chance of these 'balancing out' than in a cell without any dominating respondents.

To give a more direct visual indication of the difference in conditional coefficients of variation (CCVs) between sensitive and non-sensitive cells we show, below, boxplots of the CCVs for the different types of cells across all three industries. We also separated out marginal cells, because we would expect these to have a slightly lower amount of noise, on average, due to the larger number of respondents they contain. Also they tend to be more important cells than the interior cells, so it's interesting to see the effect on them separately.

The boxplot shows the CCVs for nonsensitive interior cells; nonsensitive marginal cells (note – all the marginal cells were non-sensitive in the tables we looked at); and the three sensitive interior cells.

Chart 2

Boxplots of the Conditional Coefficients of Variation for Different Types of Cells



All cells (88)
 Average 3.3%
 Median 2.8%
 Maximum 11%
 Minimum 0.46

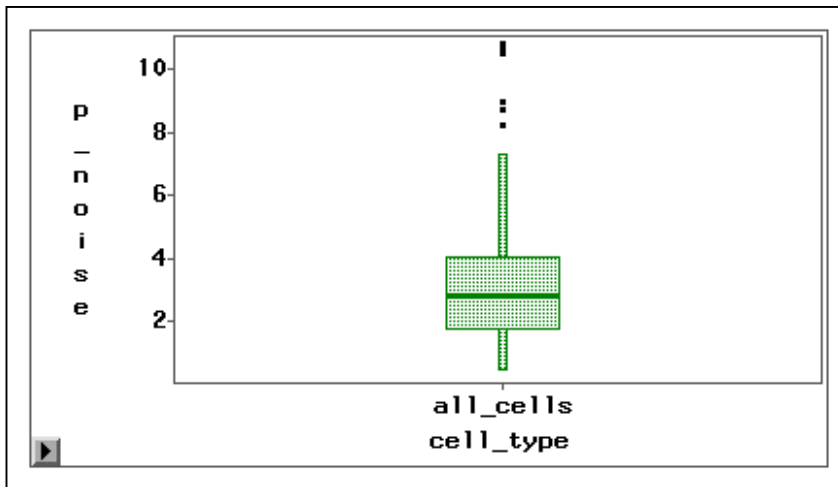
As expected, there is a significant difference between the amount of noise added to the sensitive interior cells and the non-sensitive interior cells. The marginal cells can be seen to have slightly less noise again.

6.3 The average absolute percentage noise, distributed across different types of cells

As part of our investigation of the noise method, we wanted to observe whether more noise gets added to sensitive cells and less to non-sensitive cells. We computed the absolute percentage noise in each cell for each replication. We then averaged these absolute percentages across all 1000 replications. The absolute percentage was calculated because otherwise the results would have averaged to the expected value of 0. We looked at the distribution of this 'average absolute percentage noise' across cells of different types.

Chart 3

Average Absolute Percentage Noise All Cells



Interior cells (59)

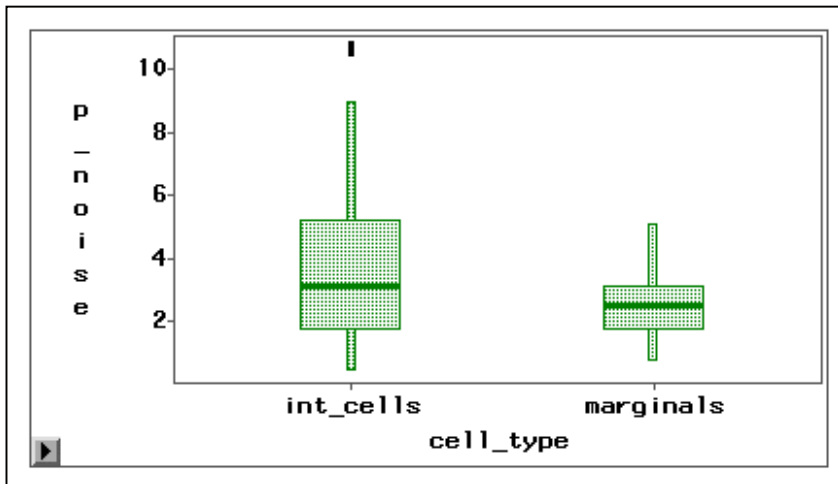
Av 3.7, Med 3.2, Max 11, Min 0.46

Marginal cells (29)

Av 2.5, Med 2.5, Max 5.1, Min 0.76

Chart 4

Average Absolute Percentage Noise Interior Versus Marginal Cells



Primary suppressions (3)

Av 10.7, Med 10.6, Max 10.8, Min 10.5

Secondary suppressions (13)

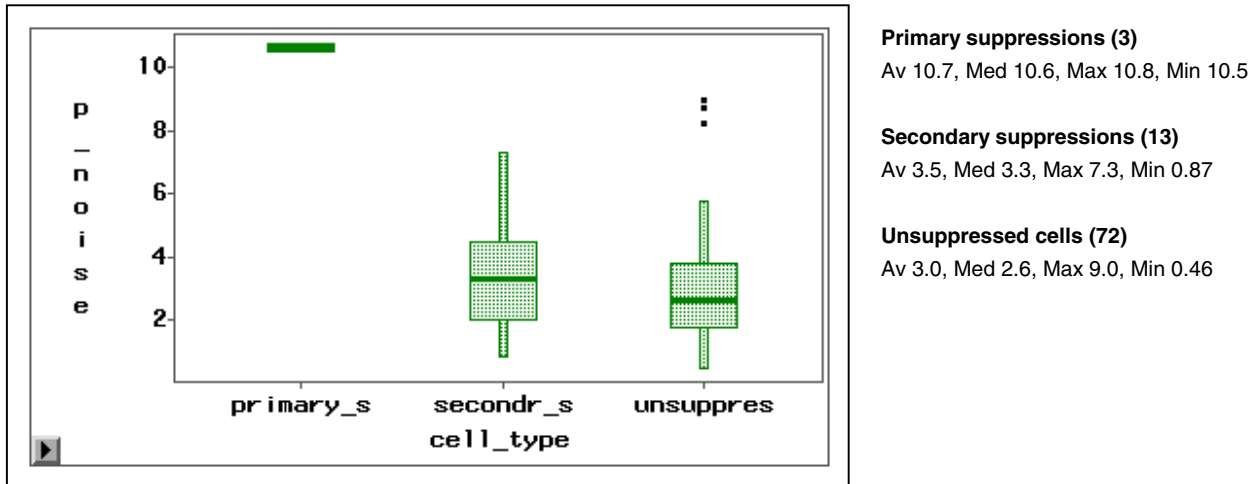
Av 3.5, Med 3.3, Max 7.3, Min 0.87

Unsuppressed cells (72)

Av 3.0, Med 2.6, Max 9.0, Min 0.46

Chart 5

Average Absolute Percentage Noise Primary and Secondary Suppressed Cells Versus Unsuppressed Cells



Marginal cells, on average, receive less noise than interior cells. This is as expected because marginal cells will generally have more contributors than interior cells. The more contributors to a cell, the more the noise in the resultant aggregate of noised values will 'balance out' and approach zero.

Primary suppressions are those cells which are defined as sensitive by the (n,k) rule. Again, and 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 is a desirable result, as these are the cells we don't need to protect against disclosure. We want the information content of these cells to be as high as possible, that is, we want the noised value to be as close as possible to the original cell value. Or, in other words, we want the total variance (due to both sampling error and confidentiality noise) to be as small as possible.

As with the conditional coefficient of variation results in Tables 8 to 10 above, secondary suppressions have slightly higher noise on average than the unsuppressed cells. Given the manual and non-optimal determination of suppression patterns, we would not really expect there to be much difference between these cells and the other non-sensitive cells. Under an optimal and/or automated cell suppression procedure, it would be reasonable to expect that the constraint of minimising the information loss would generally result in secondary suppressions of cells with small values (if total value was the measure of information being maximised). This secondary suppression would generally correspond to cells with few non-zero responses. These cells could therefore be expected to have received slightly higher noise than non-sensitive cells in general. There is a bit of this effect here, but it appears not to be significant.

While, in general, the sensitive cells will receive more noise and the non-sensitive cells will receive less noise, there is no guarantee of this. The important feature of a confidentiality method is that the uncertainty is such that an intruder knows they can never estimate a respondent's contribution with any confidence. In this case they know they can't estimate any individual respondent's value to closer than 10%.

A positive side-effect of the noise method is that those cells which will tend to have much less sampling error – cells which are mainly full-coverage or dominated by a few large full-coverage units – are those which will have more noise due to confidentialising. Conversely, those cells with more sampling error will tend to have less confidentialising noise. That is, the variance due to noise is targeted at those cells with less variance due to sampling, and away from those cells which contain more variance due to sampling.

6.4 Counts of cells having large versus small amounts of noise, for different types of cells

Table 11 shows the number of each type of cell receiving average absolute percentage noise above and below 4%. Four percent is an arbitrary threshold by which we define 'large' versus 'small' amounts of noise. But a value like this might be picked as a threshold for 'flagging'. That is, we might decide to flag cells with greater than this threshold of noise to warn users that the values may be of limited use.

Table 11

Counts of Cells According to Amounts of Noise

Large (> 4%) or Small (< 4%)

	> 4%		< 4%	
	n	Percent	n	Percent
All cells (88)	23	26	65	74
Sensitive (3)	3	100	0	0
Non-sensitive (85)	20	24	65	76
Non-sensitive (85)				
Secondary suppressions (13)	5	38	8	62
Unsuppressed (72)	15	21	57	79
Marginal cells (29)	2	7	27	93
Interior cells (56)	18	32	38	68

Table 11 repeats, in a different form, the results we have already shown above. They are that sensitive cells receive relatively more noise, and marginal cells receive relatively less noise, than the overall average amount of noise across all cells.

7. 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 added due to the noise method.

Confidentiality methods generally have to assume worst case scenarios. So, when considering the protection offered by a method such as cell suppression, we usually assume that an intruder is able and willing to perform complex calculations to derive respondents' values from the tables published. Cell suppression methods result in suppressed cells, but 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 certain level of confidentiality protection, we believe that a more realistic scenario (involving a 'non-intruding user', or 'genuine researcher') is also important (possibly more important) when trying to quantify information loss. We therefore compute two different information loss measures, corresponding to both the 'intruder' and 'non-intruder' scenarios.

1. We calculate the information loss corresponding to the feasibility intervals resulting from the particular cell-suppression pattern that was used for AES99. That is, if a user is prepared to put in the time to derive intervals for the suppressed cells in order to have approximate values rather than no values at all, this is the information loss that they would still be faced with. Given the sophistication and intent required to determine feasibility intervals for suppressed cells, we'll call this the 'intruder's information loss'.
2. However, 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, the information lost due to cell-suppression can be assumed to be the full value of the cell. This is a more naive information loss measure but, for the reasons discussed above, could be argued to more accurately represent the full impact of cell suppression for the non-intruding user. We'll call this the 'user's information loss'.

7.1 The intruder's information loss from cell suppression

The Appendix gives a simple example of how feasibility intervals for suppressed cells can be calculated.

So, how should an information loss measure be calculated from a feasibility interval? One option is to derive it from the scaled difference between the original value and the feasibility interval boundaries. Using the example in the Appendix, we have an original value (call this X_o) of 300, and a feasibility interval of (50,550). Call the lower and upper boundaries of the feasibility interval F_l and F_u respectively. So, we can define the information loss IL as

$$IL = \min\left(\left|\frac{X_o - F_l}{X_o}\right|, \left|\frac{X_o - F_u}{X_o}\right|\right)$$

Substituting the values from the example in the appendix we get

$$IL = \min\left(\left|\frac{300 - 50}{300}\right|, \left|\frac{300 - 550}{300}\right|\right) = \min(0.83, 0.83) = 0.83$$

Note that, although in this example the feasibility interval is symmetric about the original value, this is not necessarily the case.

There are problems with this approach:

1. For cells with original values of very close to zero, the use of the original value as a denominator will lead to very large values of information loss, and this may not be meaningful.
2. Another problem is that, since an intruder or a genuine user doesn't know the original value, it might make more sense to calculate the information loss corresponding to the best estimate of the original value corresponding to a confidentialised table.
3. There may be the potential to use the information loss to help derive back to the original value.
4. There is no compelling reason to use a minimum, rather than a maximum, of the differences between the original value and the bounds of the feasibility interval.

This leads to the more workable, and arguably more meaningful, option of calculating the information loss as the half-width of the feasibility interval divided by the midpoint of the feasibility interval. A consequence of this option is that feasibility intervals which include zero (as do the sensitive cells in the three AES tables we consider here) will always have the maximum information loss measure of 100%.

This option will result in the same measure of information loss (0.83) as for the example in the Appendix, because the original value happens to be equal to the midpoint of the feasibility interval.

Using this second option, we calculated the 'intruder's information loss' for the three AES99 tables and, from these, we then calculated the average intruder's information loss for each type of cell, to compare to the average absolute percentage noise using the noise method.

Table 12

Average Intruder's Information Loss

Comparison of Cell Suppression to the Noise Method

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

7.2 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 to get the following comparison:

Table 13

Average User's Information Loss

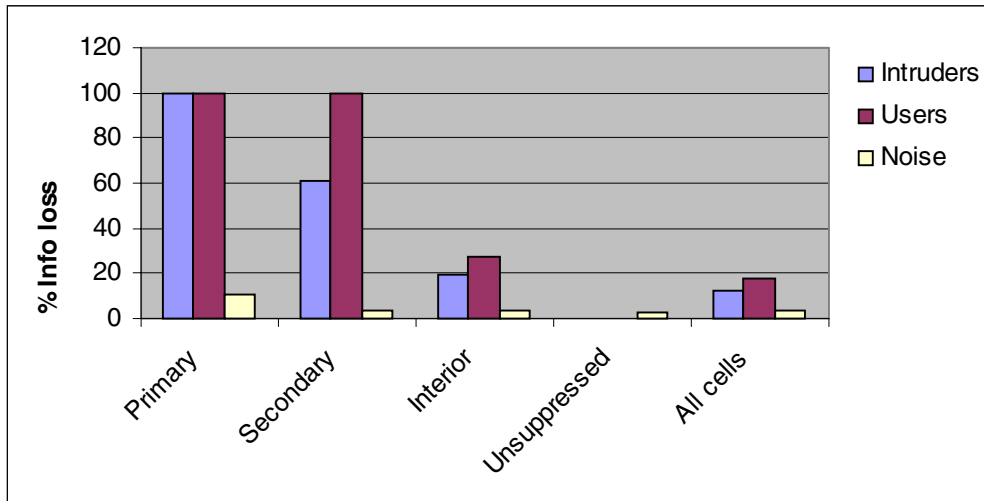
Comparison of Cell Suppression to the Noise Method

Type of Cell	Cell Suppression	Noise Method
	Percent	
Primary suppression (3)	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. The trade-off against a smaller overall information loss is the small amount of noise in the cells which would have been unsuppressed using the cell-suppression method. Figure 1 below shows the comparison visually:

Figure 1

**Comparison of Information Loss Between Cell Suppression
(For Intruders and Users) and Noise Method**



8. Future Work

8.1 Application to other outputs and surveys

The previous discussion demonstrates that the results of relative amounts of noise being applied to sensitive and non-sensitive, interior and marginal cells are as expected. However, the average level of noise for each of these types of cells will vary across variables and surveys, as it depends in part on the variance and distribution of the data which is being confidentialised. For example, the average level of noise in the three AES99 tables considered is significantly lower than that resulting from the application of the noise method to tables from the US Census Bureau's Research and Development survey, as presented in Zayatz et al (2000). This is shown by our Table 11, which achieves a similar distribution of cells with a threshold of 4%, as the US research did with a threshold of 7%. Before deciding to use the noise method to confidentialise a survey's output, it would obviously be a good idea to try it out on data from that survey, to get an idea of the level of noise to expect.

8.2 Determination of the appropriate level of noise

Further work should be put into determining the appropriate level of noise required to confidentialise data. A multiplier of around 10% has been used in this research but, as noted above, the (n,k) rule guarantees that a cell will be defined as sensitive if it contains a respondent whose value can be estimated to within $(1-k)/k$ (by a coalition of $n-1$ respondents). It could therefore be argued that an appropriate level of noise to be used in the noise method would be $(1-k)/k$ (for $k=70\%$ this would be 43%; for $k = 80\%$ this would be 25%; and for $k = 90\%$ this would be 11%). However, this assumes that cell suppression is applied optimally, and that the parameters of the (n,k) rule have been selected to assure the minimum protection required, and no more. In practice, manual selection of secondary suppressions, which doesn't ensure a minimum feasibility interval, is not likely to achieve these levels of protection against a sophisticated intruder.

8.3 Varying the distribution of the multiplier

This research has used a bimodal distribution for the multiplier, which ensures that each respondent's value is disturbed by at least 10% (while the applied function of the multiplier effectively means that other units in the sample which they are representing are left undisturbed). We could consider less stringent scenarios than this. A multiplier distribution centered on 1 with, say, a standard deviation of 10% would result in less noise, but an intruder would still have to allow for the possibility of around 10% disturbance, whether or not it has actually occurred. This is analogous to the situation of random rounding (used in Statistics New Zealand for the Population Census), where a cell has a probability of being unrounded (if it is already a multiple of the rounding base) but where the intruder doesn't know which of the values are original and which are disturbed. Using a 'unimodal' distribution would result in less information loss, but could be argued to ensure a similar level of uncertainty about the original values as is obtained from using a bimodal distribution.

8.4 Automated cell suppression

A more direct comparison of information loss from the noise method to information loss from cell suppression would use data which had been cell-suppressed using one of the automated cell-suppression packages, such as tau-ARGUS, with various optimising constraints. In this comparison, the level of uncertainty in the noise method would correspond to the parameters of the dominance rule being applied. That is, a comparison of the noise method against 'ideal' cell suppression practice would be useful when determining whether it would be a better option than cell-suppression as applied by a cell-suppression package.

8.5 Extension to count data

Tables of frequencies, or counts, can be seen as a special case of magnitude data, with every respondent having a magnitude of 1. Currently Statistics New Zealand uses random rounding (to base 3) to protect tables of counts from Population Census data. So, effectively, cell counts can be 'disturbed' by an integer value of up to 2 in either direction. By generating random unbiased additive factors taking possible values of -1, 0 and 1, we could similarly effect an unbiased integer perturbation of the cell counts. We are exploring this idea for a project which is currently being undertaken in Statistics New Zealand's datalab.

All possible three-way tables of Population Census data, corresponding to around 30 key variables, are being created as an alternative to microdata from the census. This poses problems for the application of random rounding, as the repetition of the same two-way marginal counts many times may allow derivation of the original counts through an examination of the distribution of the rounded values. While there are some ways of getting around this problem and still using random rounding, they will be relatively resource intensive, so the noise method is being considered as a confidentialising method for this project.

8.6 Consider whether flagging cells is necessary

Zayatz et al (2000) have suggested flagging or suppressing cells which are either sensitive or receive greater than a given threshold of noise (note that, in general, most cells which are sensitive will receive above the specified threshold of noise – but not necessarily). This could be reconsidered, particularly if the expected levels of noise are fairly low (as in, for example, the AES99 data examined in this paper). It may be sufficient to calculate the variance expected due to the noise, and to publish this in conjunction with information about the sampling error as a total (estimable) error.

8.7 Consider whether parameters should be kept confidential

It is standard practice to keep the parameters of confidentiality procedures (for example, the 'n' and 'k' of the (n,k) rule) confidential themselves, and this has also been suggested for the noise method. That is, the exact details of the distribution of the multiplier would be kept confidential. However, this may not be necessary. It would be useful, when explaining the method to users, to be able to give the exact details of the procedure if required and, if sufficient protection exists with the parameters known, then this may be a preferable situation in terms of user confidence in the method, and ability to meaningfully analyse the data.

9. Conclusions

We applied the noise method to AES99 data and got very similar results to earlier work by the US Census (although the average level of noise was significantly less, which is presumably a consequence of the AES dataset being larger and more homogeneous than that of the US Census Bureau's Research and Development survey). That is, we showed that, as expected, sensitive cells received significantly more noise and marginal cells slightly less noise, than the overall average level.

The noise method is a very simple method. It would be easy to apply and understand both by Statistics New Zealand staff and external users who access data through Statistics New Zealand's datalab, and are required to confidentialise their output to Statistics New Zealand standards.

For each dataset, the noise method only needs to be applied once at the microdata level and, from then on, all tables produced would be consistent with one another. This is a desirable feature, both for appearance's sake, and because it removes the risk of disclosure by linking independently confidentialised tables.

If it is possible to also protect count data using a modification of the noise method, the potential for applying one general method to all tabular output is very appealing, particularly in terms of educating users about the methods. It may be possible to have both a magnitude multiplier and a count multiplier on each dataset, to be used where appropriate (although the possible derivation of one type of data using the other would need to be explored further).

The results of this research also showed that the overall amount of noise resulting in the tables protected by the noise method compares very favourably in terms of information loss, with tables protected by cell-suppression. More work needs to be done on determining the appropriate level of noise to apply at the microdata level, and this will probably require consultation with users about how close an estimate of their value would be considered a 'disclosure'.

The introduction of this method would require careful management in terms of explaining the method and justifying the addition of noise to the data, particularly given the amount of effort that goes into designing samples which minimise variance due to sampling! It needs to be explained how information is currently being lost through the application of cell suppression to business surveys. However, the incorporation of the noise method seems no larger a task than that of random rounding, which has been successfully applied by Statistics New Zealand for some time now.

References

- Evans T, Zayatz L and Slanta T (1998). "Using noise for disclosure limitation of establishment survey", *Journal of Official Statistics*, 4:4.
- Federal Committee on Statistical Methodology (1994). *Report on statistical disclosure limitation methodology*, Statistical Policy Working Paper 22, US Office of Management and Budget, Washington, DC.
- Giessing S (2001). *New tools for cell suppression in tau-ARGUS: One piece of the CASC Project work draft*, Eurostat Work Session on Statistical Data Confidentiality, Skopje.
- 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.
- Pang S (2001). *Confidentiality issues in AES99 and Proposed recommendations to AES2000*, Statistics New Zealand Internal Paper.
- Schlackis D (1993). *Manual on disclosure control methods*, Eurostat Report, Luxembourg.
- Willenborg L and de Waal T (1996). *Statistical disclosure control in practice*, Springer Verlag, New York.
- Willenborg L and de Waal T (2000). *Elements of statistical disclosure control*, Springer Verlag, New York.
- 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

Computing Feasibility Intervals

Computing feasibility intervals can be done manually by combining the equations implied by the table. For example, consider the following fictional table:

Table A1

Fictional Unconfidentialised Table

Industry	Sales	Interest	Govt Funding	Total Inc
A	500	300	250	1050
B	750	450	600	1800
C	300	300	250	850
Total	1550	1050	1100	3700

If the cell for Industry A – interest – is determined to be a sensitive cell, and the secondary suppressions are determined on the basis of minimising the total value suppressed (a common criteria for manual suppression), the resulting confidentialised table will look like this (where ‘px’ denotes a primary suppression, and ‘sx’ a secondary suppression):

Table A2

Fictional Table with Cell Suppressions Applied

Industry	Sales	Interest	Govt Funding	Total Inc
A	500	p1	s1	1050
B	750	450	600	1800
C	300	s2	s3	850
Total	1550	1050	1100	3700

To calculate the feasibility interval for the primary suppression P1 we can use the following relationships from the confidentialised table:

- 1) $p1 + s1 = 1050 - 500 = 550$
- 2) $s1 + s3 = 1100 - 600 = 500$
- 3) $p1 + s2 = 1050 - 450 = 600$
- 4) $s2 + s3 = 850 - 300 = 550$
- 5) $p1, s1, s2$ and $s3$ are all ≥ 0

- 1) and 5) $\implies p1 = 550 - s1 \implies p1 \leq 550$
 - 1), 2) and 5) $\implies p1 = 550 - (500 - s3) = 50 + s3 \implies p1 \geq 50$
 - 3) and 5) $\implies p1 = 600 - s2 \implies p1 \leq 600$
 - 3), 4) and 5) $\implies p1 = 600 - (550 - s3) = 50 + s3 \implies p1 \geq 50$
- \implies the feasibility interval for p1 is [50,550]

Feasibility intervals for each of s1, s2 and s3 can be similarly calculated.