

Official Statistics Research The Noise Method

Summary

Confidentiality at Statistics New Zealand is a high priority issue. Continual research is carried out to find new confidentiality methods that can be applied to make more data available, and new methods that make data easier to produce. In tabular magnitude output, one of the more time consuming aspects is applying cell suppression to sensitive cells.

A possible alternative to cell suppression is applying “noise” to the underlying data. Statistics New Zealand is assessing the appropriateness of the application of noise to unit record data so that sensitive data can be included in summary statistics without breaching confidentiality. The main advantage is that once the data is noised initially, no further confidentiality protection is required.

The high level analysis indicates that a more sophisticated approach to using noised data is needed, that the underlying methodology cannot be ignored. Ignoring this aspect would lead to erroneous conclusions drawn from the data, especially where time series data is involved.

Without further investigation on how to incorporate the noise in the analysis, the noise method is not currently suitable for use in Statistics New Zealand's regular outputs, and more work is recommended.

There is also potential for using it where there is only a one-off survey, or a release that does not contain a temporal aspect. The level of noise in that case can be set high enough to provide protection without adverse results.

Contents

1. Introduction
2. Background
3. The Noise Method
4. Benefits of the Noise Method
5. Noise Method: Utility versus Confidentiality
6. Noise Creation
7. User Expectation of Noise
8. User Information about Noise
9. Analysis
10. Conclusions

Official Statistics Research, through Statistics New Zealand, commissioned this document. However, the opinions, findings, recommendations and conclusions expressed in it are those of the author(s), do not necessarily represent Statistics New Zealand and should not be reported as those of Statistics New Zealand. Statistics New Zealand takes no responsibility for any omissions, errors in, or the correctness of, the information contained in this document.

Introduction

Statistical agencies collect data from individual organizations which are often collated and used to produce summary statistics at various levels of geographic and/or industry segmentation, in particular tabular magnitude outputs. In some cases the number of data points underlying a particular cell can be quite small, particularly in areas and/or industries in which there are only a small number of players, or only a small number of large players. In such cases it may be possible to derive or accurately estimate firm level data from summary statistics, and this clearly compromises the confidentiality of the information provided by those respondents. More information about confidentiality issues in tabular data can be found in Chapter 4 of the report by the Federal Committee on Statistical Methodology.

To date, cells that Statistics New Zealand considers to be a threat to confidentiality have simply been suppressed. Cell suppression involves hiding all of the sensitive cells as well as other non-confidential cells so that basic arithmetic cannot be used to calculate the values of the sensitive cells. The method of cell suppression has created problems for users who require continuous time series data and/or complete cross-sections to perform their analyses.

A possible alternative to cell suppression is applying “noise” to the underlying data. Statistics New Zealand is assessing the appropriateness of the application of noise to unit record data so that sensitive data can be included in summary statistics without breaching confidentiality. The main advantage is that once the data is noised initially, no further confidentiality protection is required.

The noise method has been trialled before at Statistics New Zealand, based on initial work by Zayatz et al. Each project leveraged off previous work as well as greater understanding on the method being applied. For various reasons, the noise method was not ultimately used. This project is to look at the noise method and to assess its viability for protecting tabular output.

This report describes the noise methodology and discusses its relative merits, drawing on reports compiled by COVEC. The focus of this report is more on using noised data in a naïve analysis (i.e. analytical methods uninformed about the noise process) and the data utility.

The project team consisted of James Enright (Statistics New Zealand), Shane Vuletich (COVEC) and David Law (Treasury).

The Noise Method

The noise method is a way of providing confidentiality for tabular magnitude outputs by perturbing data at a microdata level. The two common methods for protection are either aggregation or cell suppression, each with their own benefits and drawbacks. The table below summarises each method by three points, which helps provide the context of how the noise method works versus the other two methods.

	Aggregation	Cell Suppression	Noise Method
Unperturbed data(*)	Yes	Yes	
Low level of data		Yes	Yes
All cells have values	Yes		Yes

It is implemented by weight perturbation. Each weight is altered so that at least a certain absolute percentage of error is built into a unit's contribution. The relative percentage of error depends on the initial weight, such that units with high survey weights do not receive a lot of relative noise error. The direction of the noise (whether increase or decrease) is held constant for a unit, although the precise amount can change on a regular basis. In the case of full coverage surveys, each unit starts with a weight of one.

The intention is such that once noise has been applied, the tables are produced with all cells safe, in that no one contributor's value can be estimated to within a certain degree guaranteed by the noise method. No further confidentialisation is necessary (although quality checks on the level of noise should be performed with possible data suppression for quality reasons).

The noise method is meant primarily for tabular magnitude output. It is possible to use it for other types of output, although other protection methods may be more suitable.

The next sections of the paper expand on the details of the noise method.

Need for the Noise Method

The increasing availability of fine level data and increasing demand for segmented and/or customized outputs have made it necessary to:

- strengthen protection to avoid the possibility of disclosure of individual data;
- implement a method of protection that is not affected by the size or the complexity of the datasets;
- use a protection system that is simple and automatic, where there are no disclosure risks posed from the multiple production of the same table or production of related tables;
- evolve confidentiality protection in New Zealand to an internationally recognized standard;

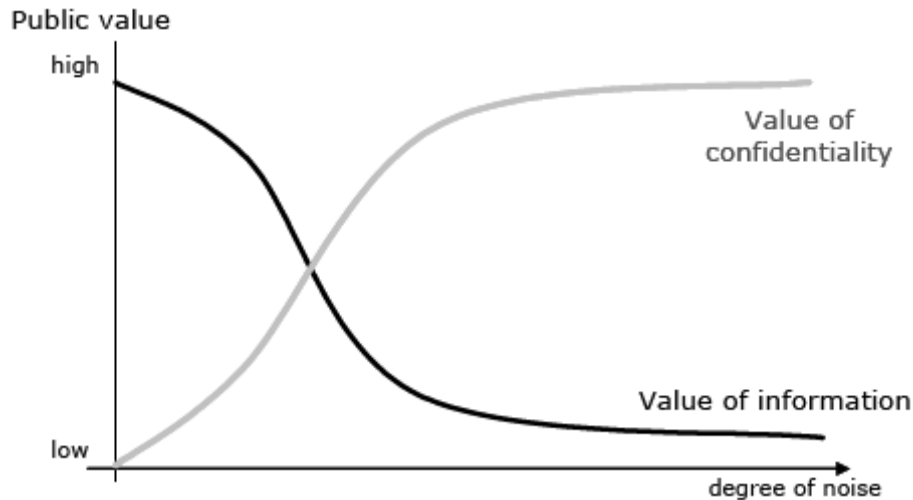
The cells with the highest disclosure risk are generally those derived from only a small number of observations. This is because a person that has some knowledge of the underlying data (e.g. someone who has contributed unit data) can use their knowledge to estimate the unit responses of the remaining contributors to that cell. In the extreme case of only two contributors each contributor would know exactly what data the other supplied simply by subtracting their own data from the cell total. Furthermore, if a business/individual represents a very large share of the market, it would be relatively easy for someone with a good knowledge of the industry to derive confidential figures.

The current cell-suppression method guarantees confidentiality by hiding the value of sensitive cells and by applying a consequential suppression to non-sensitive values. Data providers like Statistics New Zealand realize that provision of valid data to legitimate researchers delivers direct benefits to society, so a method that ensures a better balance between confidentiality and usefulness of the data is preferable.

Chapter 7 of [Krsinich and Piesse] shows that the noise method compares very favourably against cell suppression in terms of the average amount of information lost, especially because it doesn't significantly change non-sensitive cells. Once noise has been applied to data and it has been published analysts must be aware that those numbers represent approximate values only, rather than exact values. The degree of accuracy will depend on the sensitivity of the data and the grade of noise that is applied. Nevertheless, data that contains noise is would be more useful than data that includes many suppressed cells.

Noise Method: Utility versus Confidentiality

Noise can be applied to data in varying degrees depending on the sensitivity of the data. The greater the noise applied to the data, the more protected the true values will be. It is reasonable to assume that there is a public benefit associated with protecting confidential information, not only because Statistics New Zealand is legally obliged to do so but also because respondents may refuse to disclose private information or provide inaccurate data if the confidentiality of their data cannot be assured. It is also reasonable to assume that there is a public benefit in having reliable information that can be used to base important public and private sector decisions upon.



The main problem is that the 'value of confidentiality' objective is enhanced by greater noise while the 'value of information' objective is enhanced by less noise, in which case there are benefits and costs associated with the application of noise. This implies that there is some 'optimal' level of noise which balances the benefits of confidentiality against the costs of information loss.

The degree of noise that maximises the value to the public is the one which achieves the best balance between protecting confidentiality and preserving the accuracy of the information. In theory this will be achieved by applying the minimal amount of noise required to protect sensitive information.

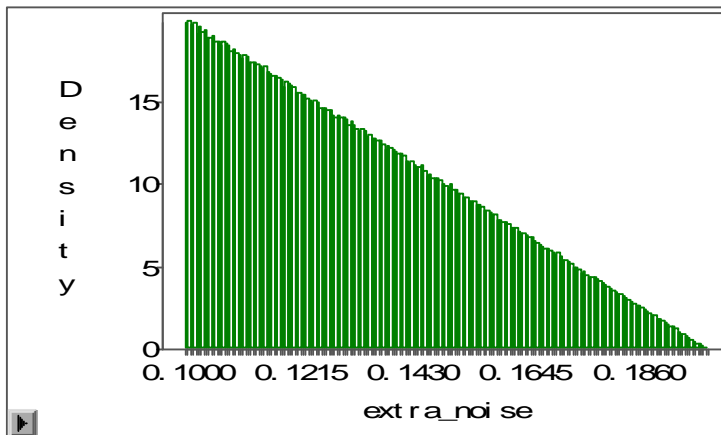
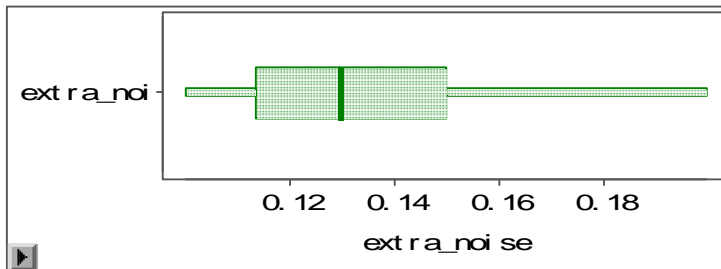
This paper is focusing more on the 'value of information' aspect, determining what degree of noise allows basic information to be useable.

Noise Creation

There are three factors to consider when applying the noise method: What is the minimal level of noise that should be applied? What is the maximal level of noise that should be applied? What is the shape of the noise distribution?

Noise Distribution

Although several noise distribution shapes have been tested before, for this project the Ramp distribution of the LEHD project¹ was used [Abowd et al]. This distribution is good in that it ensures that most of the noise is concentrated at the minimal end, and that the noise distribution behaves in a predictable fashion. The figure below shows the distribution.



Moments			
N	2553526.00	Sum Vgts	2553526.00
Mean	0.1333	Sum	340477.431
Std Dev	0.0236	Variance	0.0006
Skewness	0.5667	Kurtosis	-0.5966
USS	46815.8682	CSS	1417.9044
CV	17.6728	Std Mean	1.475E-05

Quantiles			
100%Max	0.2000	99.0%	0.1900
75%Q3	0.1500	97.5%	0.1842
50%Med	0.1293	95.0%	0.1777
25%Q1	0.1134	90.0%	0.1684
0%Min	0.1000	10.0%	0.1051
Range	0.1000	5.0%	0.1025
Q3-Q1	0.0366	2.5%	0.1013
Mode	.	1.0%	0.1005

¹ The United States 'Longitudinal Employer-Household Dynamics' project, which pioneered the noise method.

Noise Application

Noise is very simple to apply. For each unit in the data, a direction is randomly selected with equal probability so that their values are always increased or decreased. The amount of noise is randomly generated from the distribution above. From this a multiplier is created that is centered around 1 so that each value has equal probability of being decreased (multiplier<1) or increased (multiplier>1). A minimum distance from 1 is determined in percentage terms, so that a sufficient minimum disturbance is always ensured.

Only some observations deserve high confidentiality protection, and the simplest way to determine whether they belong to this category or not is their sampling weight. For example:

- A weight equal to 1 means that the individual represents itself in the survey, so its responses need to be confidentialised;
- A weight equal to 100 means that the individual is part of a large population so there's no disclosure risk.

From this logic comes the following formula:

$$\text{Data value with noise} = \text{Original data value} * (\text{multiplier} + (\text{sampling weight} - 1))$$

Since the “multiplier” is actually added to the sampling weight, its net effect will be noticeable only when the weight is low, so that non-sensitive cells are left almost undisturbed.

The constraint that sets multipliers away from 1 by a given percentage is necessary because it never allows a respondent's actual value to be exposed. Nevertheless there is always a small chance that a noised cell's value (only if made of more than one respondent) approximately equals the original one, when random multipliers manage to offset each other.

The Amount of Noise

For the minimal level of noise, ideally it should be set to such a level that if a sensitivity measure (in the case of Statistics New Zealand this is the p% rule) was applied, a cell with only one contributor would be safe. However, that much noise would likely be too much. This does mean that if noise is set below this level, it is still possible for cells to be sensitive. While, theoretically, an intruder would never know how much noise we put in and so any cell would be safe, if a respondent saw their own distorted value, they might not be too happy. The decision about what to do with still sensitive cells was left until a suitable level of noise (from an inference view point) was determined.

An initial level of noise was set so that if it was suitable, then aside from being an appropriate level of noise, it would also hopefully increase the number of non-sensitive cells. The amount of noise can be put into "Grades", where Grade 1 is the minimal amount of noise that would be usable, and there is no maximum Grade as ever more noise can be used. The initial level of noise was Grade 4.

Confidentiality: Too Little Noise

From the view of confidentiality, the noise method appeals as, ideally, once the noise method has been applied nothing further needs to be done for protection. This, however, is predicated on the idea that there is sufficient noise so that every respondent is protected. The idea is that respondents should, at the least, be protected so that an intruder can't get to within a certain percentage.

Grades 3 and 4 of noise were established so that they would provide this level of protection. This does mean that if Grades 3 and 4 were "too noisy", then the noise method cannot provide sufficient noise to be used for regular confidentialised outputs. This does not mean, though, that the noise method could not be used at all.

As such, even if Grades 3 and 4 of the noise method are "too noisy", Grades 1 and 2 will be tested to determine their impact on the ability to make inferences to inform possible future use of the noise method.

Inferences: Too much noise

From a protection perspective, the more noise the better. However, this needs to be balanced against the utility of the resulting data. In particular, at what point do inferences drawn from noisy data significantly diverge from inferences drawn from the original data?

To this end, Shane Vuletich from COVEC and David Law from Treasury provided advice about how the data in tabular magnitude outputs could be used. The first question was "is the level of noise too much?"; with the next question being, if the noise is fine at a higher level, "at what level of output does the noise become too great?". These questions are necessary to determine the practical limits of the noise method.

Shane Vuletich looked at deviations from actual values, cross sectional shares and temporal changes. Shane Vuletich had this to say about the importance of temporal changes: 'I think temporal consistency is reasonably important as we are often required to compare growth

rates. Industry groups will often want to know how their growth compared with others. I think it would be OK if the noise preserved the order of the growth rates even if it did distort the actual growth rates a little. That way we could still say with confidence that "A grew by more than B" even if we are not totally confident in the magnitude of the growth.'

User Expectation of Noise

The objective of the noise method is to protect tables against the disclosure of individuals' responses. Since the tables are often used by consultants, whose intent is not to derive confidential information but to perform analysis and draw conclusions from them, the noise method should not undermine the usefulness of the data. The intended use of the data determines how significantly the application of noise will impact on the inferences made. From a researchers perspective less noise is obviously better.

While the so-called intruder is more interested in obtaining the actual values underlying aggregated cells, the analyst is generally more concerned with the relativities between cross-sectional shares and/or temporal changes. The analyst is much more likely to use the published data to make comparisons, rank items or calculate ratios. The table below summarizes the ways in which noise could adversely affect some of the things an analyst might like to do with the data.

INTENDED USE OF THE DATA	VALUES	CROSS-SECTIONAL SHARES		TEMPORAL CHANGES		
	Deviation from Actual Values	Deviation from Actual Shares	Alteration of the order	Deviation from Actual Changes	Directional consistency	Alteration of the order
Use with other external data	X			X	X	
Comparison across industries		X	X			
Time series analysis				X	X	
Ranking of items			X			X
Calculation of ratios	X		X			

Effect on Values

The actual value is by definition the one that needs to be masked whenever a confidentiality risk occurs. In a magnitude table, when an aggregated total includes a small number of respondents and/or dominant individuals, the multiplier that is applied should be sufficiently large to prevent a skilled intruder from deriving or accurately estimating the original values. From an analytical perspective the resulting deviations should not compromise the usefulness of 'noised' figures when combined with other data sources. Furthermore, if two or more noised values are used to calculate a ratio, the result could be significantly distorted if the respective noise scalars go in opposite directions.

Effect on Cross-Sectional Shares

A cross-sectional share is derived by expressing the value of an individual cell (component) as a percentage of all of the components within a specified class within a single time period. For example, an analyst may want to know the percentage of total retail sales accounted for by supermarkets.

In this case the supermarket is the component and the specified class is total retail sales which is made up of the sales of all individual retail components (including supermarket sales). The cross-sectional share for supermarkets is derived by dividing supermarket sales by total retail sales in the selected time period.

The share that each cell represents within a certain industry can be distorted by the noise according to its strength and direction. Cross-sectional shares do not represent sensitive information per se but they can be important for analytical purposes. It is not easy to define

the size of the deviation that would not affect the conclusions of an analysis, because it depends on how the values are distributed. For example, if two or more actual values are sufficiently close to each other, even a relatively small noise could reverse their order.

Effect on Temporal Changes

Temporal changes are derived by calculating the percentage changes in individual cells or groups of individual cells over time. For example, an analyst may want to know how supermarket sales have changed over time, or how total retail sales have changed over time. The temporal changes in supermarket sales are the percentage changes in supermarket sales between each time period.

Since the noise method involves the application of a random multiplier to each observation, it is expected that the values within a given category will have different distortions applied to them in different years. Depending on the strength of the noise applied, this has potentially serious implications for time series analysis and forecasting and, if sub-annual figures are available, on seasonal analysis. For example, when actual temporal changes are small, the noise is potentially capable of reversing the direction of those changes, leading the analyst to erroneous conclusions: this is particularly dangerous when only a short time period (i.e. 3-4 years) is taken into account. Over longer periods, the noise is much more likely to balance itself out, so the results of the analysis would be approximately preserved.

User Information about Noise

When deciding to apply a new confidentiality procedure, a statistical agency should be aware that a certain degree of transparency is important for users. The publication of data tables is usually accompanied by details about the sampling methodology, the sample error and any adjustments that are made to the data (e.g. random rounding). COVEC believes that the application of noise should also be disclosed to users of the data.

Cell Flagging

While cell suppression prevents users from viewing or analyzing some data, it ensures that all of the visible information is unperturbed. In contrast, tables containing noise have no suppressed cells but the downside is that it is not possible to determine which cells are accurate and which are not.

COVEC therefore thinks that it would be advisable to flag cells that have had significant amounts of noise applied to them. This would allow analysts to understand the limitations of the data and assess the overall reliability of their work e.g. in terms of the number of high distortion cells included in the calculations. The obvious question is what the flag threshold should be set at. Alternatively (or additionally), knowing the variance generated by the noise would be highly desirable since it would give users a numerical indication of the accuracy of the tables. The variance would be larger for cells with smaller numbers of observations or smaller numbers of dominant respondents.

Confidential Parameters

Transparency through cell flagging and/or variance disclosures will increase user confidence but it may also allow intruders to get approximate values of confidential data. The explanation by the statistical agency may include:

- The formula that adds the multiplier to the sample weight;
- The distribution of the multiplier;
- The minimum and maximum noise applied in percentage terms;

While COVEC doesn't think that cell flagging is a threat to confidentiality, some problems may arise if variances and/or details about how the noise is applied are disclosed because these details would allow skilled intruders to compute feasibility intervals for confidentialised data. In order to preserve its original intent the statistical agency would therefore need to apply a considerable amount of noise to the data which would undermine the value of the data significantly. Lack of information can generate disturbance in its own right, without actually decreasing the reliability of the dataset.

Information from Agencies

Statistics New Zealand agrees with COVEC that information about the noise method should be provided, although the exact parameters used should be kept confidential. Also, highly perturbed cells should be flagged to enable a proper indication of data quality.

Analysis

What was analysed

The data provided for analysis was based on the Annual Enterprise Survey (AES) outputs. High level tables from AES were recreated using noised data, and the results inferred from the noised data were compared to the results inferred from the original data.

The analysis of the noised data was carried out by Shane Vuletich of COVEC. For this investigation, only a small sub-set of data was analysed, namely the *Total Expenditure* sub-category of the *All Industries* table. Also, attention was focussed on the 2004 data.

Although this is a small sample of output to examine, this is from highly aggregated data, where little noise is expected and no cells are sensitive, so no comparison to cell suppression is possible. If analysis at this level is problematical, this indicates that lower levels would perform worse.

The following components were analysed:

- Deviation from actual values. This is simply the noised value minus the original value.
- Cross-Sectional Shares. This looks at the relative contributions of the components to the total expenditure.
- Temporal Changes. This looks at the movements from 2003 to 2004.
- Directional Consistency. This looks at the direction of the changes between years.
- Order of Cross-Sectional Shares. This looks at the ranks of the cross-sectional shares.
- Order of Temporal Changes. This looks at the ranks of the temporal changes.

Grades 1 to 4

The noise was initially set at Grade 4. Based on the results, this was lowered through Grade 3, Grade 2 and finally set to Grade 1. The combined results are presented below.

Grades 3 and 4 represented a sufficient amount of noise from the view of Statistics New Zealand, although there were still possibilities of more protection being required. Grades 1 and 2 were considered to provide information on potential use of noise in other situations (such as for CURFs).

In all cases, the amount of noise was found to be too high to allow clear analysis of temporal changes and temporal rankings.

Figure 1 contains the graphical summaries of the variations from actual values for each Grade. The deviations were found to be acceptable for all Grades.

Figure 2 contains the graphical summaries of the variations in the cross-sectional shares for each Grade. The deviations were found to be acceptable for all Grades.

Figure 3 contains the graphical summaries of the variations in the temporal changes for each Grade. The deviations were found to be acceptable for all Grades. From the report: "The noise

has altered the estimated growth rates between 2003 and 2004. Some of the changes are quite significant and could be misleading if reported as actual changes. The deviations from actual do not seem to be strongly related to the grade of noise applied.”

Figure 1 Deviation from Actual Values in 2004

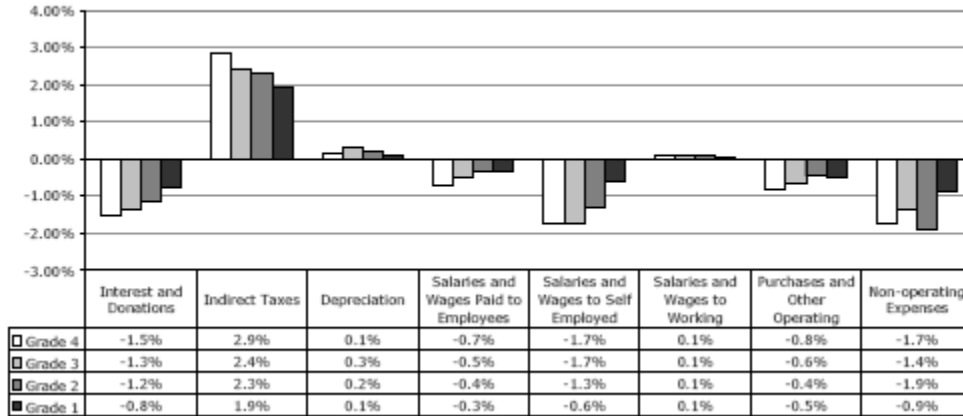


Figure 2 Cross Sectional Shares in 2004

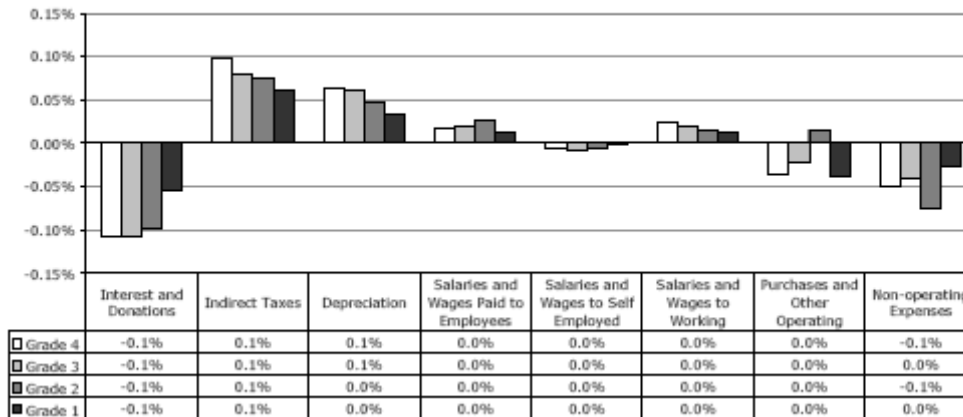


Figure 3 Temporal Changes between 2003 and 2004

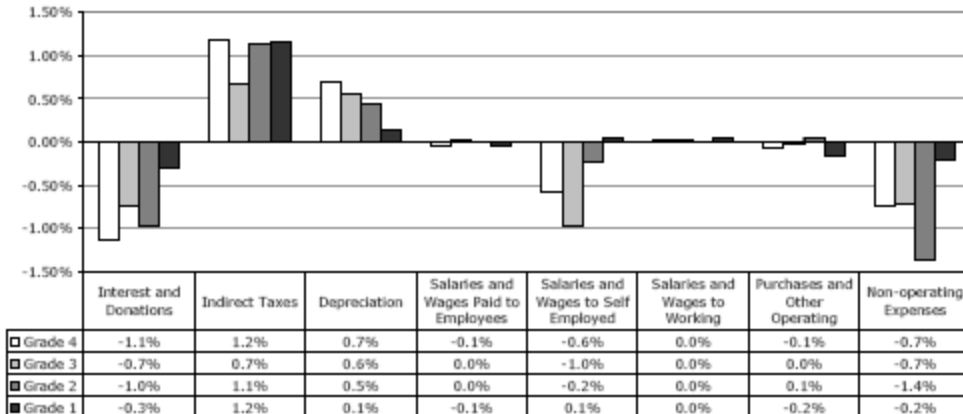


Table 1 contains the summaries of the directional consistency for each Grade for the temporal changes. The directions of the changes were consistent for all Grades.

Table 2 contains the summaries of the ranks of cross-sectional values for each Grade. Where ranks changed due to noise, the original ranks were close in value, and thus these changes were not deemed to be a major issue.

Table 3 contains the summaries of the ranks of temporal changes for each Grade. The changes here are of a larger order than for the cross-sectional shares and were deemed to adversely impact on the temporal analysis.

Table 1 Directional Consistency

Expenditure Category	Grade 4	Grade 3	Grade 2	Grade 1
Interest and Donations	Same	Same	Same	Same
Indirect Taxes	Same	Same	Same	Same
Depreciation	Same	Same	Same	Same
Salaries and Wages Paid to Employees	Same	Same	Same	Same
Salaries and Wages to Self Employed Commission Agents	Same	Same	Same	Same
Salaries and Wages to Working Proprietors	Same	Same	Same	Same
Purchases and Other Operating Expenses	Same	Same	Same	Same
Non-operating Expenses	Same	Same	Same	Same

Table 2 Order of Cross-Sectional Shares 2004 (Ranked by Value)

Expenditure Category	Grade 4	Grade 3	Grade 2	Grade 1
Interest and Donations	Same	Same	Same	Same
Indirect Taxes	Different	Different	Different	Different
Depreciation	Same	Same	Same	Same
Salaries and Wages Paid to Employees	Same	Same	Same	Same
Salaries and Wages to Self Employed Commission Agents	Same	Same	Same	Same
Salaries and Wages to Working Proprietors	Different	Different	Different	Different
Purchases and Other Operating Expenses	Same	Same	Same	Same
Non-operating Expenses	Same	Same	Same	Same

Table 3 Order of Temporal Changes 2003-04 (Ranked by Value)

Expenditure Category	Grade 4	Grade 3	Grade 2	Grade 1
Interest and Donations	Different	Same	Different	Same
Indirect Taxes	Same	Same	Same	Same
Depreciation	Same	Same	Same	Same
Salaries and Wages Paid to Employees	Different	Different	Different	Different
Salaries and Wages to Self Employed Commission Agents	Same	Same	Same	Same
Salaries and Wages to Working Proprietors	Same	Same	Same	Same
Purchases and Other Operating Expenses	Same	Same	Same	Same
Non-operating Expenses	Different	Different	Different	Different

COVEC reports that the amount of noise is too disruptive for temporal analysis. Given that the level of noise involved is less than desired by Statistics New Zealand, the conclusion from this is that the noise method cannot be used where there is temporal analysis. Since the majority of Statistics New Zealand's outputs have significant temporal components, this implies that the noise method is not suitable as a standard confidentiality tool.

The next section of the project considered the cross-sectional data, which was considered useable under Grades 1 to 4.

Grades 7, 15, 27

Given that under Grades 1 to 4, the cross-sectional data was useable, the question was raised as to what Grade of noise made them unusable?

Grade 7 was selected as a more comfortable level of noise from a confidentiality perspective. Grade 15 is a significant amount of noise, but still preserves some original information. The noise in Grade 27 can destroy all original data, but was tried for interest.

Figures 1 to 3 show the deviation in actual values, cross-sectional shares and temporal changes. Grade 7 produces acceptable deviation in cross-sectional data, but Grades 15 and 27 are introducing too much deviation. The cross-sectional shares can still be analysed under Grades 7 and 15, but not under Grade 27. All Grades negatively impact on temporal changes (as expected).

Tables 1 to 3 show the directional consistency of temporal changes and the rankings of cross-sectional and temporal changes. In this regard, Grades 7, 15 and 27 are the same as other Grades in that they are directionally consistent, preserve cross-sectional orderings to within acceptable uses but distort temporal rankings.

Figure 1 Deviation from Actual Values in 2004

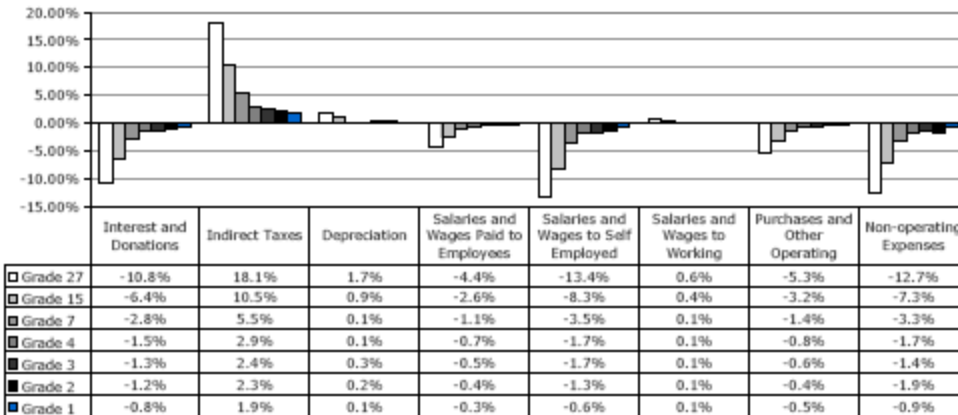


Figure 2 Cross Sectional Shares in 2004

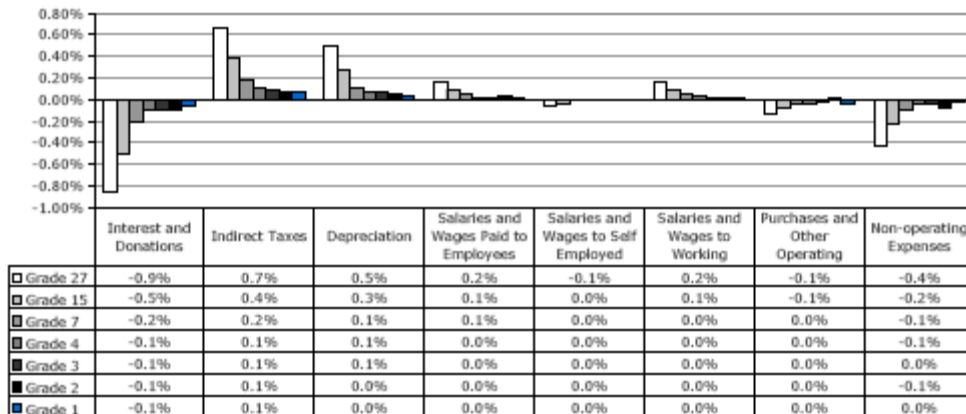


Figure 3 Temporal Changes between 2003 and 2004

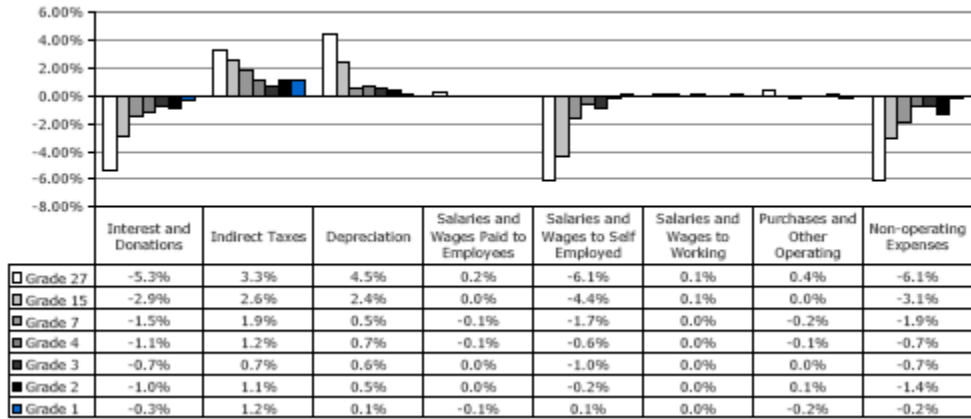


Table 1 Directional Consistency

Expenditure Category	Grade 27	Grade 15	Grade 7	Grade 4	Grade 3	Grade 2	Grade 1
Interest and Donations	Same	Same	Same	Same	Same	Same	Same
Indirect Taxes	Same	Same	Same	Same	Same	Same	Same
Depreciation	Same	Same	Same	Same	Same	Same	Same
Salaries and Wages Paid to Employees	Same	Same	Same	Same	Same	Same	Same
Salaries and Wages to Self Employed Commission Agents	Same	Same	Same	Same	Same	Same	Same
Salaries and Wages to Working Proprietors	Same	Same	Same	Same	Same	Same	Same
Purchases and Other Operating Expenses	Same	Same	Same	Same	Same	Same	Same
Non-operating Expenses	Same	Same	Same	Same	Same	Same	Same

Table 2 Order of Cross-Sectional Shares 2004 (Ranked by Value)

Expenditure Category	Grade 27	Grade 15	Grade 7	Grade 4	Grade 3	Grade 2	Grade 1
Interest and Donations	Same	Same	Same	Same	Same	Same	Same
Indirect Taxes	Different	Different	Different	Different	Different	Different	Different
Depreciation	Same	Same	Same	Same	Same	Same	Same
Salaries and Wages Paid to Employees	Same	Same	Same	Same	Same	Same	Same
Salaries and Wages to Self Employed Commission Agents	Same	Same	Same	Same	Same	Same	Same
Salaries and Wages to Working Proprietors	Different	Different	Different	Different	Different	Different	Different
Purchases and Other Operating Expenses	Same	Same	Same	Same	Same	Same	Same
Non-operating Expenses	Same	Same	Same	Same	Same	Same	Same

Table 3 Order of Temporal Changes 2003-04 (Ranked by Value)

Expenditure Category	Grade 27	Grade 15	Grade 7	Grade 4	Grade 3	Grade 2	Grade 1
Interest and Donations	Different	Same	Different	Different	Same	Different	Same
Indirect Taxes	Different	Same	Different	Same	Same	Same	Same
Depreciation	Different	Different	Different	Same	Same	Same	Same
Salaries and Wages Paid to Employees	Different	Different	Different	Different	Different	Different	Different
Salaries and Wages to Self Employed Commission Agents	Different	Different	Different	Same	Same	Same	Same
Salaries and Wages to Working Proprietors	Same	Same	Same	Same	Same	Same	Same
Purchases and Other Operating Expenses	Different	Different	Same	Same	Same	Same	Same
Non-operating Expenses	Different	Different	Different	Different	Different	Different	Different

The impact on the cross-sectional analysis is a problem under Grade 27, and even under Grade 15. Grade 7 is a comfortable level of noise, sufficient for confidentiality protection and hence indicates that the noise method can be used as a tool for protecting data that has no temporal aspect.

Conclusions

From the high level analysis it is clear that a naïve approach to analysing noised data would lead to erroneous conclusions, especially in the case of time series data. This means that currently the noise method cannot be used by Statistics New Zealand in regular production without understanding more about how to use the output.

No comparison of the inference of noised data to the inference from suppressed data can be drawn from this analysis as there was no suppressed data to examine.

There is also potential for using it where there is only a one-off survey, or a release that does not contain a temporal aspect (possibly a tabular release or a CURF). The level of noise in that case can be set high enough to provide protection without adverse results (a Grade in the 10 to 20 area would provide enough protection while not impacting significantly on high level results).

Extra Work

In this report, the focus was on whether or not the noise method was appropriate for typical Statistics New Zealand outputs. Further work, a potential OS Research project, could focus on how to incorporate information about the noise method into analysis, how does information loss under noise compare to information loss under cell suppress, and about complex analysis on noised data.

Work Not Addressed

In the project plan, more work was planned on analysing noised data. The following areas were to be considered, and are still open questions:

1. Given that the amount of noise is unknown: is a noised cell still sensitive if the un-noised cell was?
2. What level of noise in a cell renders the cell unusable? (This depends on the type of analysis.)
3. Is noise restricted to magnitude tabular output, or can it be applied to other outputs, such as tables of counts?
4. What information about noise can be released for researcher's use?
5. Can the application of noise be incorporated in analysis (as opposed to naively using noised data)?

References

Abowd, John M., Stephens, Bryce E. and Vilhuber, Lars, "Confidentiality Protection in the Census Bureau's Quarterly Workforce Indicators", LEHD website, 2005

Federal Committee on Statistical Methodology, "Statistical Disclosure Limitation Methodology", Working Paper 22, Office of Management and Budget, 2005 (revised)

Krsinich, Frances and Piesse, Andrea, "Multiplicative Microdata Noise for Confidentialising Tables of Business Data", Statistics New Zealand, Jan 2002

Zayatz, Laura, Evans, Timothy and Slanta, John, "Using Noise For Disclosure Limitation of Establishment Data", US Bureau of the Census, 2000