Confidentialising Business Demography output tables using the Noise for Counts and Magnitudes (NCM) method

November 2016

Contact: frances.krsinich@stats.govt.nz

1. Overview

Statistics NZ has introduced an 'input perturbation' approach to confidentialising business demography tables called the 'Noise for counts and magnitudes' (NCM) method.

Input perturbation involves adding a small amount of 'noise' to the data at the unit (ie business or person) level, in such a way that the tables derived from this perturbed data are unbiased and contain as much information as possible while targeting protection to the sensitive cells.

Input perturbation is used by other statistical agencies

Perturbation methods are being used in production by a number of other official statistical agencies. In particular, the US Census Bureau uses a 'noise infusion' method to protect longitudinal employment data (Abowd *et al*, 2012), and the Australian Bureau of Statistics use noise in the protection of frequency tables accessed via their remote server TableBuilder (Chipperfield *et al*, 2016).

This approach was first considered by Statistics New Zealand for the case of business survey magnitude tables (Krsinich and Piesse, 2002) with an application to count data researched by Groom and Camden (2013).

Business Demography Statistics

Business Demography statistics are derived from Statistics NZ's Business Register and comprise of statistics based on two of the Registers' statistical units, the enterprise (ENT) commonly known as the business and the geographic unit (GEO) commonly known as the business location. It produces a longitudinal series of statistics based business counts (counts of enterprises and geographic units) and business employee counts (employee counts of the enterprise or geographic unit) by a variety of business and regional classifications

A coordinated approach

We have developed an approach which perturbs both count and magnitude tables – the NCM method - and this is being considered more widely across the organisation as part of the development of an automated confidentiality service.

Note that, in the context of Business Demography, the respondent whose confidentiality is being protected is the business. This means that tables of employee counts are considered magnitude tables, as the number of employees is a magnitude with respect to the business.

How it works

Each business is assigned a random number uniformly distributed between 0 and 1. This random number is fixed across time to ensure the same degree of perturbation is applied to the business over time.

See section 2 for an example of the basic NCM approach, and section 3 for details of how the NCM method is adapted for the specific case of Business Demography data.

The benefits

The benefits of this NCM method compared to the previous confidentialisation method are that:

1. there will be more data released, and
2. related tables will be consistent with each other – that is, the same cell in related tables will have the same value.

2. The NCM method

Business counts

For count tables, the business-level random numbers are used to generate a new random number for businesses grouped together in a cell, and this is the basis for a 'fixed' version of random rounding to base 3 (FRR3) which will ensure that the same group of businesses will always be rounded the same way in related tables.

Employee counts

The random number is used to generate a 'noise multiplier' which feeds into the generation of magnitude tables (ie employee counts).

Cells at risk of disclosure are those with either a small number of contributors, or a few very large contributors. For these cells there will be relatively more noise, while for the cells with no disclosure risk the noise tends to cancel out at the cell level.

Individual values are protected by at least +/- 10% so, for the most vulnerable cells with only one business, we guarantee this level of uncertainty about the employee count of that business. For cells composed of many businesses the noise will tend to cancel out. We can flag cells with more than a certain level of noise so that analysts can treat these values with caution. We will consult on what threshold value of cell-noise should be flagged.

Example of the basic NCM method

Here is an example of the basic NCM method, without the extra features that have been incorporated for the particular case of the BD data.

Underlying data

Consider the following set of businesses, belonging to ANZSICs (industry) A, B and C and regions Auckland and Wellington. Each business (identified by a GEO number) is assigned a 'random seed' which is a uniformly distributed random number between 0 and 1.

GEO nbr	ANZSIC	Region	Employee count	random seed
g01	A	Auckland	120	0.047
g02	B	Auckland	54	0.377
g03	B	Auckland	2	0.988
g04	C	Auckland	7	0.640
g05	C	Wellington	33	0.035
g06	B	Auckland	54	0.746
g07	B	Wellington	187	0.422
g08	A	Wellington	166	0.630
g09	B	Auckland	350	0.819

g10	C	Auckland	32	0.118
g11	A	Auckland	9	0.510
g12	A	Wellington	8	0.959
g13	C	Auckland	47	0.111
g14	C	Wellington	50	0.457
g15	B	Wellington	42	0.964

Sorted by ANZSIC and region

We sort the data by ANZSIC and Region to make the derivation of cell totals easier to follow in the next steps of the example:

GEO nbr	ANZSIC	Region	Employee count	random seed
g01	A	Auckland	120	0.047
g11	A	Auckland	9	0.510
g08	A	Wellington	166	0.630
g12	A	Wellington	8	0.959
g02	B	Auckland	54	0.377
g03	B	Auckland	2	0.988
g06	B	Auckland	54	0.746
g09	B	Auckland	350	0.819
g07	B	Wellington	187	0.422
g15	B	Wellington	42	0.964
g04	C	Auckland	7	0.640
g10	C	Auckland	32	0.118
g13	C	Auckland	47	0.111
g05	C	Wellington	33	0.035
g14	C	Wellington	50	0.457

Noised employee counts

To ensure that the noised data is unbiased we need equal odds of adding or subtracting a small amount of noise at the business level. In this example we add or subtract 10% noise with equal probabilities of 50%.

Note that in the application of the method to Business Demography data, there is a small amount of extra noise added[1] but for the purposes of this example we consider the simpler approach of adding or subtracting exactly 10%.

Also, for Business Demography data, we treat employee counts of under 10 differently by adding or subtracting 1 – this is detailed in section 3 under the subsection 'Noising employment counts'. Again, for the purposes of this example, we consider the simplest case of adding or subtracting 10% to all employee counts.

So, if the random seed is less than 0.5 we multiply the employee count by 0.9 and if the random seed is greater than 0.5 we multiply the employee count by 1.1.

[1] See the appendix for details of how this extra noise is derived.

In our example, g01 has its employee count of 120 multiplied by 0.9 to get a noised employee count of 108.00 because the random seed of 0.047 is less than 0.5.

On the other hand, g11's random seed is 0.510, which is greater than 0.5. So the noised employee count for g11 is 9 x 1.1 = 9.90.

GEO nbr	ANZSIC	Region	Employee count	random seed	noised employee count
g01	A	Auckland	120	0.047	108.00
g11	A	Auckland	9	0.510	9.90
g08	A	Wellington	166	0.630	182.60
g12	A	Wellington	8	0.959	8.80
g02	B	Auckland	54	0.377	48.60
g03	B	Auckland	2	0.988	2.20
g06	B	Auckland	54	0.746	59.40
g09	B	Auckland	350	0.819	385.00
g07	B	Wellington	187	0.422	168.30
g15	B	Wellington	42	0.964	46.20
g04	C	Auckland	7	0.640	7.70
g10	C	Auckland	32	0.118	28.80
g13	C	Auckland	47	0.111	42.30
g05	C	Wellington	33	0.035	29.70
g14	C	Wellington	50	0.457	45.00

Tables of employee counts – before and after 'noising'

Original employee counts

	Auck	Wgtn	
A	129	174	**303**
B	460	229	**689**
C	86	83	**169**
	675	**486**	**1161**

Noised employee counts

	Auck	Wgtn	
A	117.90	191.40	**309.30**
B	495.20	214.50	**709.70**
C	78.80	74.70	**153.50**
	691.90	**480.60**	**1172.50**

Percentage difference between original and noised employee counts

	Auck	Wgtn	
A	-8.60	10.00	2.08
B	7.65	-6.33	3.00
C	-8.37	-10.00	-9.17
	2.50	-1.11	0.99

Fixed random rounded to base 3 (FRR3) business counts

Original counts of businesses

	Auck	Wgtn	
A	2	2	**4**
B	4	2	**6**
C	3	2	**5**
	9	**6**	**15**

Derivation of the cell-level random seeds

To derive the cell-level random seeds, we sum the random seeds of the businesses in the cell and apply the modulo 1 function. In other words, we extract the non-integer part of the sum of the business's random seeds.

So, for example, the sum of the random seeds of the four businesses in ANZSIC B and Auckland is 0.377+0.988+0.746+0.819 = 2.931. The cell-level random seed is the non-integer part of this: 0.931.

This cell-level random seed is also uniformly distributed between 0 and 1, which enables direct application of the random rounding.

Sum of the random seeds

	Auck	**Wgtn**	
A	0.558	1.589	**2.146**
B	2.931	1.385	**4.316**
C	0.870	0.492	**1.362**
	4.358	**3.466**	**7.824**

Cell-level random seeds

	Auck	**Wgtn**	
A	0.558	0.589	**0.146**
B	0.931	0.385	**0.316**
C	0.870	0.492	**0.362**
	0.358	**0.466**	**0.824**

Random rounding of the business counts

The random rounding of each cell is based on the cell-level random seeds. The random rounding is 'fixed' in the sense that a cell with a given set of contributors will always be rounded the same way across different tables because the cell-level random seed is determined by the sum of the contributors' random seeds.

Random rounding to base three involves leaving all multiples of three unchanged. Counts which are not already a multiple of three have a 2/3 probability of being rounded to the nearest multiple of three, and a 1/3 probability of being rounded to the further multiple of three.

This is operationalised as follows:
If the count is not already a multiple of three and the cell-level random seed is less than 2/3 (or 0.667) then we round to the nearest multiple of three. Otherwise, if the cell-level random seed is greater than 0.667 we round to the further multiple of three.

Original counts

	Auck	Wgtn	
A	2	2	**4**
B	4	2	**6**
C	3	2	**5**
	9	**6**	**15**

Cell-level random seeds

	Auck	Wgtn	
A	0.558	0.589	**0.146**
B	0.931	0.385	**0.316**
C	0.870	0.492	**0.362**
	0.358	**0.466**	**0.824**

FRR3 counts

	Auck	Wgtn	
A	3	3	**3**
B	6	3	**6**
C	3	3	**6**
	9	**6**	**15**

So, the original count of 3 for ANZSIC C and Auckland is 3, which is left as 3 in the FRR3 table since it is already a multiple of 3.

The original count of businesses in ANZSIC A and Auckland is 2. The cell-level random seed is 0.558, which is less than 0.667 so we round this to the nearest multiple of 3, which is 3.

There are 4 businesses in ANZSIC B and Auckland, and this cell has a random seed of 0.931, which is greater than 0.667, so we round this to the further multiple of 3 which is 6.

Note that the marginal cells – ie total counts for ANZSICs, regions, and the total table count of 15, are random-rounded independently of the internal cells, rather than derived from the internal cells.

This is why the total FRR3 business count for Auckland is 9 rather than 12 (ie the total of the Auckland internal cells, 3+6+3).

The non-additivity of the marginals is actually a positive feature of the FRR3 method if suppression of small cells in sparse tables is required[2] because there will then be no secondary suppression required to protect against derivation from marginal totals.

3. The NCM method applied to Business Demography data

Derivation of the random seeds

Longitudinal random seeds

The data is released for the full time series from 2001 to 2016, so we set the random seeds for each statistical unit (ENT or GEO) for the full longitudinal record. That is, a GEO whose employment counts are being perturbed by 2% in 2016 will have also had their employment counts perturbed by 2% in 2015.

Break in the longitudinal random seed setting between 2014 and 2015

Until 2014, the Business Demography data was released unconfidentialised. Therefore there would be some disclosure risk associated with having the same random seed at the business level across the entire longitudinal record from 2000 to 2016, as the original and noised data between 2014 and 2001 can be compared.

Because of this, we break the longitudinal random seed between 2014 and 2015. That is, there is a longitudinal random seed assigned at the business level from 2000 to 2014, and then a new random seed is assigned from 2015 onwards. A consequence of this is that there will be more noise in the changes at cell-level between 2014 and 2015 than between other adjacent years.

Coordinating the random seeds for geographic units and enterprises

The seeds are derived from the random number stored on the Business Register at the geographic unit (GEO) level.

We coordinate the ENT seeds with the GEO seeds so that corresponding tables will be perturbed as similarly as possible. For single-GEO ENTS, the ENT random seed is the same as the GEO random seeds. For multi-GEO ENTS the ENT random seed is set the same as the first GEO's random seed (where the GEOs are sorted according to the geographic unit number).

For enterprises belonging to a group of enterprises under common ownership (referred to as GTE Level in the Business Register) we reset the random seeds at the enterprise level so that all enterprises under common ownership have the same random seed. This is done to ensure that we are protecting the GTE-level information to the required level (ie at least 10%).

Fixed random rounding to base 3 for counts of businesses

Tables of counts of businesses are produced alongside tables of cell-level random seeds[3], and the cell-level random seeds are the basis for the fixed random rounding to base 3 (FRR3).

---

[2] See subsection 'suppression of small counts not required' in section 3 for an explanation of why suppression can be required in other contexts, such as population census tables.

[3] We take mod1 of the sum of the random seeds associated with the contributors to the cell – this is itself a uniformly distributed random number between 0 and 1.

In addition, we add further uncertainty to cells with rounded counts of 0 by rounding original counts of 3 down to 0 with a 1/3 probability, up to 6 with a 1/3 probability, and leaving them as is with a 1/3 probability.

The reason for this extra step is that there can be FRR3 cells of 0 businesses corresponding to non-zero employee counts. Without the addition rounding of original 3's, there is a disclosure risk posed by knowing the that business count must be 1 or 2. With the additional rule, a FRR3 cell of 0 with a corresponding non-zero employment count corresponds to a business count of either 1,2 or 3.

Noising employment counts

The usual approach to input perturbation of magnitude tables is to add a certain percentage of noise to the magnitude. In the case of small employment counts, though, this might not provide sufficient protection.

For example, consider a business with 3 employees. Adding 10% noise to this employee count would give us 3.3 employees, and if this is the only business contributing to the cell then that count is rounded to the nearest integer in the graduated rounding stage, which in this case would make the noised value also 3.

To deal with this issue we have added an extra step to the perturbation of employee counts. For (non-zero) original counts of less than 10 we subtract 1 from the count with a probability of 1/3, we add 1 with a probability of 1/3 and we leave the count as is with a probability of 1/3.

For counts of 10 and more we use the standard approach of adding or subtracting 10% (plus some extra noise) with probability of ½ in either direction.

Graduated rounding of noised employment counts

For original counts of 10 and over, the noising procedure will usually result in non-integer values, so we apply a further rounding step, with increasing rounding bases as the noised employee count gets bigger. This ensures both that the employee count is an integer, and that there is not a greater level of precision implied in the cell count than is warranted.

For noised employee counts under 22 we round to 3; for counts from 22 and under 100 we round to 5; for counts from 100 and under 1000 we round to 10; for counts from 1000 and under 5000 we round to 50; and for counts 5000 and over we round to 100.

Employee size group statistics

Removing tables of employee counts in employee size groups

Previously, we have released both business counts and employment counts within employee size groups. There is a disclosure risk associated with this, however, as employment counts can be used in some cases to derive the original business counts for the corresponding size range.

While the release of employee counts within employee size groups is useful for many users, note that the same information can be estimated from the business count tables where required (by multiplying the number of businesses by the mid-point of the employee size group).

For future years we will consult with users about whether distributional information would be useful to more directly target the analytical needs formerly being met by the employee counts in employee size groups.

Restricting the classification level for which employment size range business counts are released

The level of classifications for which counts of businesses in different employment size ranges are released will be restricted to high levels of the classfication, as there is a disclosure risk at fine levels of classification if there is only one employment size range represented in the data.

Finer classification level employee size group tables will be available as customised requests where as assessment of the disclosure risk can be made on a case-by-case basis. Another option is for users to work with the data at finer levels within the protected Datalab environment.

Suppression of small counts not required

In some other contexts, such as the population census, random rounding to base 3 (RR3) is supplemented by suppression of small cells in very sparse tables. This is to avoid disclosures though inference that the few '3's are likely to have been rounded up from '1's. There is also a very small probability of RR3 counts being breakable because of particular combinations of interior and marginal RR3 cells only corresponding to particular original counts. The cell suppressions help to guard against this situation also.

For business demography tables, on the other hand, there is not a risk of disclosing new information from the tabulation variables used to define the business count tables – because this information is already in the public domain. Also, the original business counts could not be used to help derive the magnitude tables (of employee counts), which are already protected to at least 10% even when the exact corresponding counts are known.

Therefore we propose that no cell suppression is required for the business demography output tables. A final decision on this will be informed by the testing and consultation.

Zeros in business counts tables

In the business count tables which are now protected by FRR3, a 0 cell corresponds[4] to either a 1,2 or 3. Zero cells will be indicated as such by a '..' in the tables.

We have differentiated between real 0's and FRR3 0's that correspond to non-zero original values as we believe this will give more utility to users representing all the 0 cells (including those FRR3 to 0 from non-0 original cells) while still protecting confidentiality.

Indicating high levels of noise in the employment count tables

A recommended development is the flagging of employment count cells where the noised and rounded cell count is more than a certain percentage from the unconfidentialised count.

This will enable users to treat these cells with more caution when basing conclusions upon them, and to assure users that unflagged cells have a noised value which is known to be close to the original value.

We did seek user feedback on how best to flag cells with higher levels of noise.

Inter Year data protection

Because certain cells may have the same contributors from one year to the next any inter year comparison will reveal the percentage change. However as the percentage change is calculated from

[4] See the section on fixed random rounding which explains that we add further uncertainty by random rounding of counts of 3, which will result in 1/3 of 3's being rounded down to 0.

the rounded data this will ensure the actual data is protected if there are the same one or two contributors to the cell.

Percentage change statistics requested as customised jobs will be based on the rounded data.

Appendix. Algorithms used to apply the NCM method to Business Demography tables

Selected extracts of the SAS code used to apply the NCM method are shown here.

Fixed random rounding to base 3 (FRR3)

- The cell-level random seed is derived from the sum of the business's[5] random seeds.
- Original counts of 3 are rounded up to 6 or down to 0 with a 1/3 probability each, or left as is with probability 1/3.
- Other multiples of 3 are left unchanged.
- Non-multiples of 3 are rounded to the nearest multiple of 3 with probability 2/3 and to the further multiple of 3 with 1/3 probability.

```
cell_ran_seed_geo = mod(geo_random_seed,1);

 if cell_ran_seed_geo < 0.67  then do;
    if unconf_geo_count=3 and cell_ran_seed_geo <= 0.33 then frr3_geo_count = 0;
    else if mod(unconf_geo_count,3) = 1 then frr3_geo_count = unconf_geo_count-1;
    else if mod(unconf_geo_count,3) = 2 then frr3_geo_count = unconf_geo_count+1;
    else if mod(unconf_geo_count,3) = 0 then frr3_geo_count = unconf_geo_count;
 end;
 else if cell_ran_seed_geo >= 0.67 then do;
    if unconf_geo_count=3 then frr3_geo_count = 6;
    else if mod(unconf_geo_count,3) = 1 then frr3_geo_count = unconf_geo_count+2;
    else if mod(unconf_geo_count,3) = 2 then frr3_geo_count = unconf_geo_count-2;
    else if mod(unconf_geo_count,3) = 0 then frr3_geo_count = unconf_geo_count;
 end;
```

2-stage noising of employment counts

- If the original employee count is less than 10, then there is a 1/3 probability of either adding or subtracting 1 and a 1/3 probability of leaving it unchanged.
- For employee counts of 10 or more, 10% of the employee count (plus a small extra factor based on the random seed) is added or subtracted with ½ probability in either direction.

```
if geo_emp_count  = 0 then noised_geo_emp_count = geo_emp_count_nbr;

else if geo_emp_count < 10 then do;

if geo random seed <= 0.33 then sub10 = -1;
else if geo_random_seed >= 0.67 then sub10 = 1;
else if geo_random_seed > 0.33 and geo_random_seed < 0.67 then sub10 = 0;

noised_geo_emp_count = geo_emp_count + sub10; end;

else if geo emp count >= 10 then do;
if geo_random_seed < 0.5 then noise_multiplier = 0.9 - (0.5-geo_random_seed)/100;
else if geo_random_seed >= 0.5 then noise_multiplier = 1.1 + (geo_random_seed-0.5)/100;

noised geo emp count = geo emp count * noise multiplier; end;
```

[5] Either the geographic unit or the enterprise. In this case we show the code used for GEOs.

Graduated rounding

- The noised employee counts are rounded with increasing bases – to achieve integer values and so that the level of precision is not misrepresented.

```
 if noised_geo_emp_count < 22 then do;
    gr_noised_geo_emp_count = round(noised_geo_emp_count,3);
end;
else if noised_geo_emp_count < 100 then do;
    gr_noised_geo_emp_count = round(noised_geo_emp_count,5);
end;
else if noised_geo_emp_count < 1000 then do;
    gr_noised_geo_emp_count = round(noised_geo_emp_count,10);
end;
 else if noised_geo_emp_count < 5000 then do;
    gr_noised_geo_emp_count = round(noised_geo_emp_count,50);
end;
else if noised_geo_emp_count >= 5000 then do;
    gr_noised_geo_emp_count = round(noised_geo_emp_count,100);
end;
```

References

Abowd, J.M., Gittings, K., McKinney, K.L., Stephens, B.E., Vilhuber, L. & Woodcock, S. (2012, April). *Dynamically consistent noise infusion and partially synthetic data as confidentiality protection measures for related time-series*. Presented at FCSM. Available at: http://digitalcommons.ilr.cornell.edu/ldi/5/

Groom, P. and Camden, M. June 2013. *Replacement for RR3 – Standard Tool for Confidentialising Count Tables*. Internal Statistics New Zealand paper.

Chipperfield, J., Gow, D. and Loong, B. 2016. *The Australian Bureau of Statistics and releasing frequency tables via a remote server*. Statistical Journal of the IAOS 32. Available at http://content.iospress.com/articles/statistical-journal-of-the-iaos/sji969

Krsinich, F. and Piesse, A. 2002. *Multiplicative microdata noise for confidentialising tables of business data*. Statistics New Zealand technical paper. Available at: http://www.stats.govt.nz/~/media/Statistics/browse-categories/business/business-character/multiplicative-microdata-noise-bus-data/mmnconbusdata.pdf