Establishing an Automated Confidentiality Service in Stats NZ

Antony Gomez[*], Frances Krsinich[**] and Allyson Seyb[***]

[*] Stats NZ, antony.gomez@stats.govt.nz
[**] Stats NZ, frances.krsinich@stats.govt.nz
[***] Stats NZ, allyson.seyb@stats.govt.nz

Abstract: Stats NZ is in the process of developing an Automated Confidentiality Service (ACS) to meet increasing demand from customers to be more flexible in providing confidentialised data for analysis and decision-making. The service will provide self-service products while meeting the organisation's statutory requirements of maintaining privacy and confidentiality of persons and businesses. A perturbation method being investigated for business outputs and the 2018 Census is the Noise for Counts and Magnitudes (NCM) method. The NCM method is a relatively simple method that can be easily implemented and provides consistency in repeated outputs. The NCM method can be adapted to enable geospatial confidentiality for counts of persons within customised geographies. We will present our findings on the NCM method and our proof of concept trials for an ACS.

1 Introduction

Stats NZ (formerly Statistics NZ) has a vision of unleashing the power of data to change lives. Its aim is to increase the value of data and move towards a more open data environment. It has more recently taken over the role of leadership for the Open Government Information and Data Programme, encouraging and supporting government agencies, Crown organisations and local authorities to make their data more freely available. The Open Data responsibilities will also build on the Data and Analytics leadership role Stats NZ has been asked to take on in support of the Better Public Services programme.

Stats NZ is bound by the Statistics Act 1975 (NZ Government, 2013) to maintain privacy and confidentiality by not disclosing information about an individual or business. Applying confidentiality techniques to data is the main approach we use to minimise the chance of disclosure for disseminated data.

Statistics New Zealand's four values for confidentiality are:

- Data utility
- Safety in managing the risks of disclosure
- Simplicity and practicality in our methods
- Consistency across collections and modes of output.

The desired outcome when considering disclosure control methods is where data utility is maximised and the risk of disclosure is low ("safe"). Improving one of these measures will usually come at the cost of decreasing the other.

The confidentiality rules Stats NZ applies for disseminated data can found in the Stats NZ Microdata Output Guide (Stats NZ, 2016). However these rules are often applied manually and are time-consuming. To move to a more open and accessible data environment there is a real need for an Automated Confidentiality Service (ACS).

2 What is an Automated Confidentiality Service?

An ACS delivers confidentialised outputs, primarily tables, in an automated way. It is a service and not a tool though it may contain many tools. Neither is it the methodology for confidentialising data. What ACS requires is the assignment of a permanent random number [0,1] to each statistical unit.

It allows for confidentialised cell outputs to be derived consistently and accurately. It will reduce the need for manual application and checking, saving on time and resources. Requests for customised tables can be generated online directly by the customer rather than the present method of submitting a request through the Stats NZ customer service team.

This will lead to more open and accessible data increasing the value of the data, its use and the quality of research. It reduces the barrier to entry allowing for greater use by the public. It also provides an incentive for other government and non-governmental organisations to share their data with Stats NZ as part of our functional data leadership role for big data and analytics.

3 Noise for Counts and Magnitudes Method

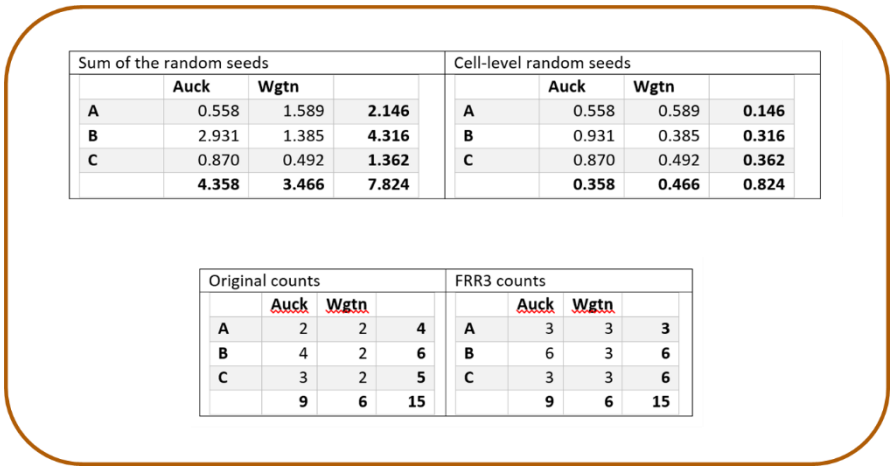
The Noise for Counts and Magnitudes Method (NCM) is two methods, one for perturbing counts or frequencies on the output side while the other deals with perturbing magnitudes or values on the input side. Both methods require a random number [0,1] to be permanently assigned to each statistical unit in the data (Krsinich & Piesse, 2002; Krsinich 2016).

3.1 Counts

Each confidentialised cell count is based on a cell-level random number which is derived by aggregating the random numbers of the contributing unit records to the cell and dropping the integer part of the sum. The cell-level random number is used to round the original cell count using the Fixed Random Rounding to base 3 (FRR3) method. If the cell-level random number is less than and equal to ⅔ the original count is rounded to its nearest multiple of 3. Greater than ⅔ and the count is rounded to its next nearest multiple of 3. This ensures, that based on probabilities, the mean is the original count. Counts that are already a multiple of 3 including 0 remain the same.

The method is simple and easy to apply. Cells with the same contributing unit records are rounded consistently even if the table structure is different. If primary suppression

is used, i.e. for low counts, the non-additivity of the marginal cells means no secondary suppression is required. In this case the output cell values are perturbed.



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**

Original counts

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

FRR3 counts

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

Figure 3.1 Example of FRR3 on counts of businesses.

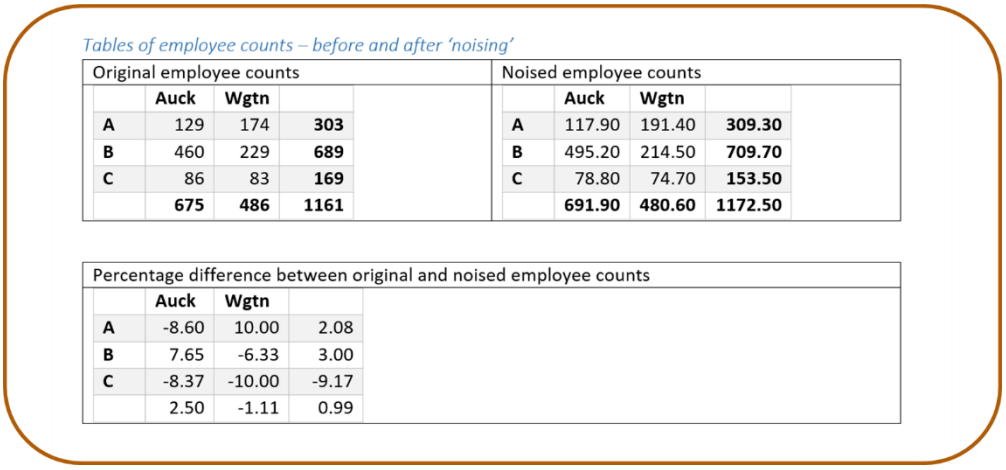
3.2 Magnitudes

Confidentialised magnitudes or values are derived using a noise multiplier at the unit record level. A small level of perturbation is specified e.g. 10%, and a noise multiplier is derived based on the unit record random number and the level of perturbation. As an example, if the random number for the unit record is less than or equal to 0.5, the noise multiplier for that unit record would be equal to 0.9 – (0.5 – random number)/100. If greater than 0.5 the noise multiplier is equal to 1.1 + (random number – 0.5)/100. The noise multiplier is used to multiply the magnitudes to derive the perturbed values. Confidentialised cell magnitudes are obtained by aggregating the perturbed magnitudes of the contributing unit records. This 'input perturbation' approach is a variant of the EZS noise method proposed by Evans, Zayatz and Slanta (1998).

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

Figure 3.2 Example of perturbed business employee counts.

Again the method is simple and easy to apply. The same cell is perturbed the same way even in a differently structured table. The noise is targeted to sensitive cells, those with low numbers of contributing records while the noise tends to cancel in cells with a large number of contributing records. Additivity is preserved for the marginal cells and no suppression is required due to the nature of the perturbation.



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

Figure 3.3 Example of perturbed magnitude cell outputs of business employee counts.

4 Project Wero

Stats NZ is working with external commercial companies to do proof of concept initiatives under Project Wero (Mäori - a challenge). Project Wero is about meeting the needs of our customers, future-proofing Stats NZ, thinking big and challenging ourselves.

One proof of concept project involves application software to interactively create insights from data using visualisation and analytical tools. The confidentialising of the output data is done on the fly rather than confidentialising the input data before using the application. This project focuses primarily on the output and visualisations to gain insights to the data.

The second proof of concept project involves delivering a confidentiality API (application programme interface) which also confidentialises data on the fly. Built into the API is a perturbation method such as the ABS Tablebuilder product (Chipperfield *et al*, 2016) or the NCM method. This project is focused on the confidentialising of data on the fly but the API can be integrated with other software including open source software such as the mapping software NationalMap for example.

Both projects are in the testing phase and if Stats NZ chooses either one or both of these systems it will form a significant part of the ACS.

5 Implementing the Automated Confidentiality Service

ACS is being implemented in a step by step process in different subject matter areas within Stats NZ. The NCM method is being tested to see whether it meets the current confidentiality requirements for the outputs produced by the subject matter areas. As Stats NZ moves to implement a new data model for its leadership in open data, the hope is that the ACS will become a single service within that data model.

5.1 Business Demography

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

The Business Demography team were the first to trial the NCM method. The respondents that need to be protected in business demography statistics are the businesses. FRR3 is applied to the business counts whereas employee counts are considered a magnitude and a noise multiplier is used at the unit record level. See Figures 3.1, 3.2 & 3.3.

The trial was largely successful in that there was less information loss than in the previous release of the business demography tables especially at the regional level. The NCM method has been built into the production process for generating the tables (Krsinich, 2016).

When developing the NCM approach for production of the business demographic tables, there was extensive user consultation. Feedback from users has been very positive.

5.2 Census 2018

The New Zealand Census of Population and Dwellings is to take place in March 2018. Most Census are tables of counts and currently the confidentiality for the dissemination of this data is based on 8 rules (consisting of identification and protection rules) that address low values, small area tables, sparseness, and derived measures. One of the key methods, random rounding to base 3 (RR3), has been in use in the organisation for a long time. Unlike FRR3, it is based on an arbitrary random seed to confidentialise cell counts. It does have some weaknesses, the main one being that under repeat application it can reveal the cell true counts.

With a need for customers to be able to create their own detailed output via an easy-to-use tool, tables need to be confidentialised automatically and have consistent cell values. To address this issue, the Census 2018 team is undertaking comparison testing between the NCM and the ABS Tablebuilder perturbation methods. This testing uses

code written in the open source programming language R. A decision to use one of these methods will be made and the code integrated as part of the Census output production process.

Geospatial confidentiality also has to be addressed for Census 2018 for use in customised area boundaries and map visualisation tools. See section 6 below.

5.3 Integrated Data Infrastructure

The Integrated Data Infrastructure (IDI) is a large secure database of linked microdata used for social research. The microdata includes Stats NZ survey data including Census, government agencies' administrative data as well as data from some non-government organisations. Access to the de-identified data is given to approved researchers in a secure Datalab environment.

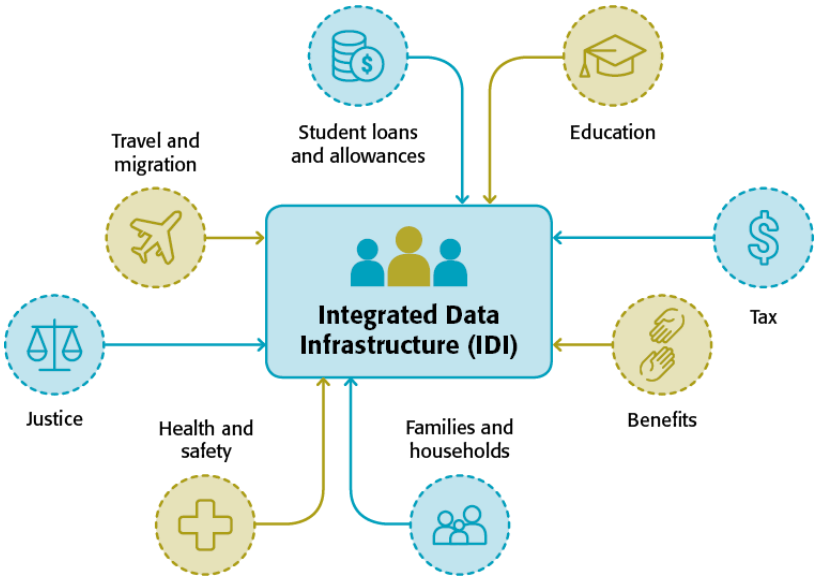


Figure 5.1 A diagrammatic view of the IDI (Stats NZ, 2017).

Outputs are confidentialised by the researchers themselves according to Stats NZ's current confidentiality rules, then checked by Stats NZ staff before they are approved for release. There are two checks; Phase 1 is to release tabular and other statistical outputs to others on the project team whereas Phase 2 is to release a report, publication or presentation for general use.

The Datalab is a difficult environment for the operation of an ACS due to varied nature of the analysis undertaken by the researchers. Basic macros are available e.g. Excel, SAS, R, to assist with confidentialisation but the researchers are required to implement these themselves. What is currently being automated is the checking process where a Python (or R) script can process a folder with multiple Excel files containing a large number of individual sheets (tables) to produce a comprehensive

diagnostic report. This saves the checker hours of time manually opening each file and checking each sheet containing a table. It is envisaged that at some stage in the future, the researchers will have access to a reliable ACS which will eliminate the need for the Phase 1 checks.

5.4 Registered-based Statistical System

At Stats NZ we are planning to develop a Register based Statistical System (RSS) (Wallgren & Wallgren, 2014). The model for such a system includes three base registers, representing people, businesses, and locations. These base registers will link together, and also link to many other sources of data related to their respective units. Links from the person register and business register to the location register is via addresses whereas the link from the person register to the business register is via the Linked Employer-Employee Data (LEED). Other sources of data will be linked to one or more of these base registers.

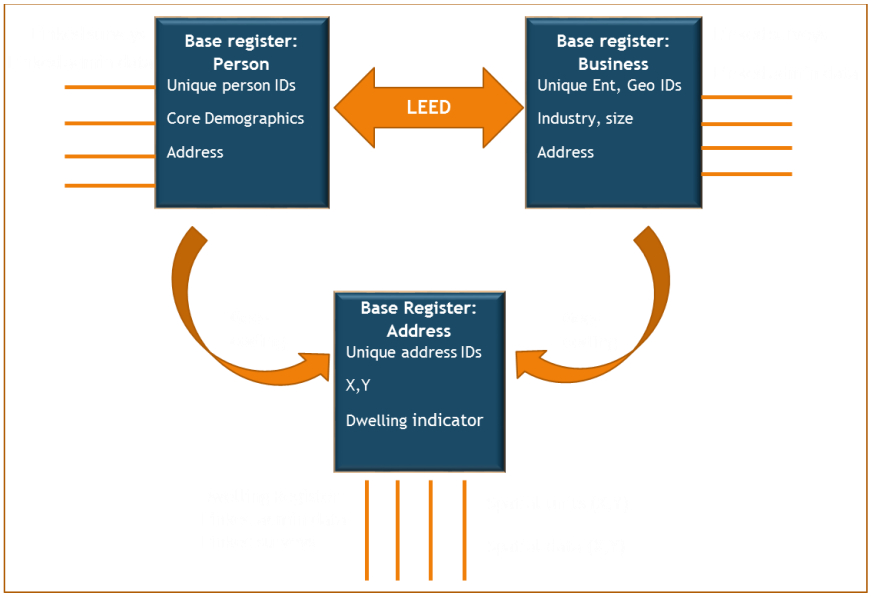


Figure 5.2 A simplified view of an integrated register-based system (Bycroft, 2016).

The three base registers are at different stages of development: a mature Business Register maintaining a list of businesses and enterprises for use in sample selection for business surveys has existed for several decades; a Statistical Location Register has been developed to maintain a list of addresses geocoded with (x,y) coordinates for use in collection operations in the Census of Population and dwellings in 2018. Still at the conceptual stage, is the person register.

To incorporate the ACS into the RSS, random numbers have to be assigned to each of the statistical units in each of the three base registers. Other linked sources of data will derive their unit record random number from the base registers based on the

statistical unit that needs to be confidentialised. The use of random numbers for an ACS is part of the RSS high level design and principles (Bycroft, 2017).

6 Geospatial Confidentiality

One of the main issues with geospatial confidentiality is dealing with counts of persons or dwellings and estimates of measures over a non-meshblock area, a meshblock being the small geographical area unit for which statistical data is collected and processed by Stats NZ. A non-meshblock area could contain a number of meshblocks but the boundary of the area of interest would cut through some of the meshblocks. Spatial enablement through location (x,y) geocoding allows for area boundaries and dwellings to be defined by an (x,y) coordinate system. The problem that arises is how to protect the identification of individual dwellings, especially if boundaries can be moved or changed that would isolate dwellings through differencing. A new proposed solution to this problem has been made possible through the development of the ACS and the NCM method where permanent random numbers are assigned to each unit record whether it be a person, dwelling or business.

6.1 Customised Geographical Areas of Interest

Meshblocks are Stats NZ's lowest geographical output level which in most cases consists of a number of dwellings and people.

In Figure 6.1 the blue squares represent the individual meshblocks with numbers of persons in each. The area of interest is shaded in green. Meshblocks is the lowest output level where counts of persons or dwellings are randomly rounded to base 3. The dwellings are geocoded to their location in the location register.

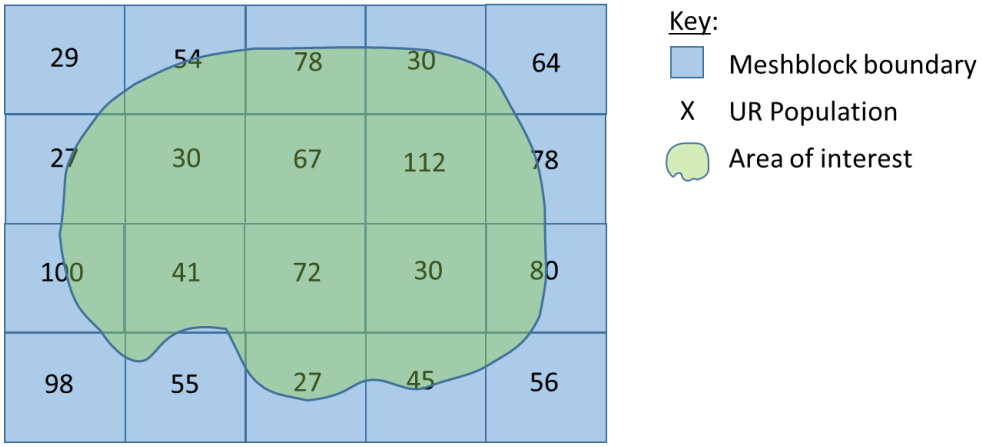


Figure 6.1 Area cutting across urban/rural (UR) meshblocks (Morgan, 2016).

The problem that arises is when boundaries are changed slightly and through differencing (hatched area), dwellings can be isolated and therefore poses a risk of disclosure as shown in Figure 6.2.

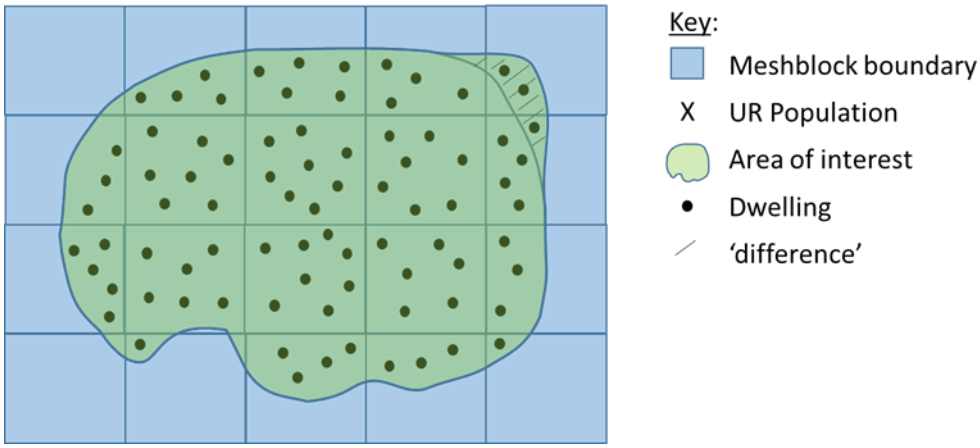


Figure 6.2 Changes to the boundary of the area of interest.

One solution that has been proposed looks at an estimation process where dwellings in split meshblocks are proportioned out to other split meshblocks based on where the area boundaries lies. This effectively leaves some meshblocks empty and others with additional dwellings. In some sense this realigns the area boundary with the meshblock boundaries while ensuring estimates within the area boundary are not biased. If the area boundary is changed then the number of dwellings within the assigned meshblocks will change but not necessarily the meshblocks themselves. If the change is substantial then it would mean adding or subtracting meshblocks to those originally selected.

Although this proposed solution to the non-meshblock geospatial confidentiality problem is workable, it appears difficult to manage and not readily adaptable for automation.

6.2 New Proposed Solution

To achieve consistency in output tables for ACS, a fixed random number (between 0 and 1) is assigned to each unit record whether it be a person, dwelling or business. This random number also provides us with a way of selecting dwellings within a split meshblock to achieve unbiased estimates while ensuring confidentiality.

As before, the proportion of dwellings inside the boundary of a split meshblock is determined by which dwellings lie inside the area of interest within each split meshblock (Figure 6.3). The random number assigned to each dwelling is used to select whether the dwelling lies inside or outside the area boundary. For example if the proportion of the number of dwellings within a split meshblock is 0.75 then all

dwellings with a random number less than (and equal to) 0.75 is selected to lie within the area of interest.

This way the selection of the dwelling is not based on the (x,y) location within the meshblock but is randomly selected to represent the meshblock. Confidentiality is preserved as it would be difficult to determine which dwelling is assigned to be within the boundary and which is assigned to lie outside without knowledge of the dwelling random number. Consistency is also preserved as it uses the same fixed random number for a dwelling every time the same output is generated.

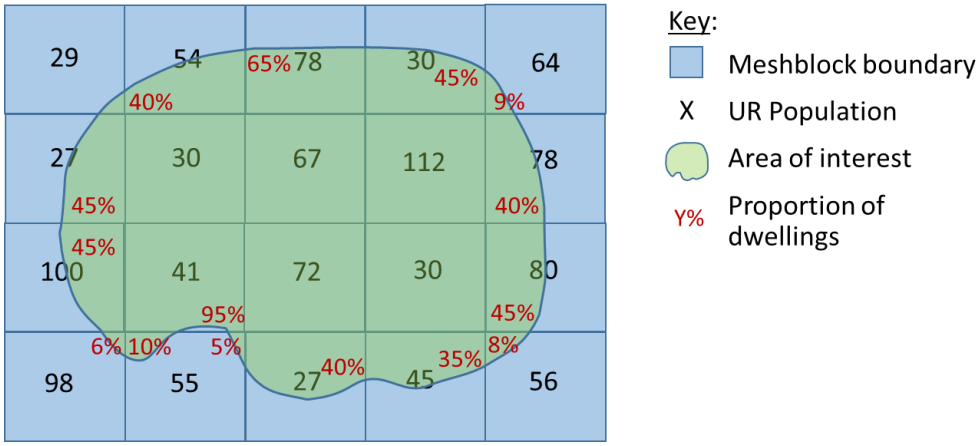


Figure 6.3 Dwelling with a random numbers less than or equal to the proportion are included within the area of interest.

Changing the area boundary may change the proportion of dwellings lying inside the area boundary within a split meshblock and this would change the number of dwellings selected. It is important to note that consistency is always preserved using this method.

Final outputs of counts of persons, dwellings or businesses would still follow confidentiality rules such as FRR3 and this can be applied using the NCM method where the fixed random numbers are used for confidentialising.

The proposed allocation of a fixed random number to each unit record is primarily there for the confidentialising process and not the selection of persons, dwellings or businesses. Using the same random number for confidentiality and selection could possibly lead to disclosure through unpicking. This implies that the use of a second independent random number for each unit record is desirable.

However the digits of a random number are in themselves random and unbiased. So we can easily use an unbiased transformation function such as multiplying the random number by 10 and dropping the integer part to determine a second unbiased random number. This number can then be used for the selection process for estimation.

In fact with enough stored digits for each random number gives the ability to generate additional random numbers to do other forms of selection or confidentiality processes. The allocation of a fixed random number to each unit record provides a way of achieving consistency not only for confidentiality but also for estimation purposes.

This allocation of a fix random number to each unit record not only confidentialises the output of counts or magnitudes but can be used to solve non-meshblock geospatial confidentiality. It achieves consistency in the output, does not need to be managed in terms of what output has been previously made public, and the method can easily be automated as part of the ACS.

7 Summary

An ACS will enable Stats NZ to meet the needs of an increasing demand from customers to be more flexible in providing confidentialised data for analysis and decision-making, while maintaining the organisation's statutory requirements for the privacy and confidentiality of persons and businesses. The introduction of random numbers at the unit record level and development of the NCM method will allow the organisation to achieve its Open Data goals through the ACS. The NCM method is simple and easy to apply, which has led to a solution for confidentialising customised geographical areas.

References

Bycroft, C. (2016). *A data model to connect location, social and economic data and statistics.* Internal Stats NZ paper.

Bycroft, C., Matheson-Dunning, N. & Seyb, A. (2017). *Design elements for a statistical person register.* Internal Stats NZ paper.

Chipperfield J., Gow, D. & Loong, B. (2016). *The Australian Bureau of Statistics and releasing frequency tables via a remote server.* Statistical Journal of the IAOS 32. http://content.iospress.com/articles/statistical-journal-of-the-iaos/sji969Z

Evans, T., Zayatz, L. & Slanta, J. (1998). *Using Noise for Disclosure Limitation of Establishment Tabular Data.* Journal of Official Statistics, Vol.14, No.4, pp. 537–551. http://www.jos.nu/Articles/abstract.asp?article=144537

Krsinich, F., & Piesse, A. (2002). *Multiplicative microdata noise for confidentialising tables of business data*. http://www.stats.govt.nz/~/media/Statistics/browse-categories/business/business-character/multiplicative-microdata-noise-bus-data/mmnconbusdata.pdf.

Krsinich, F. (2016). *Confidentialising Business Demography tables using the noise for counts and magnitudes (NCM) method*. Internal Stats NZ paper. https://www.researchgate.net/publication/311734931_Confidentialising_Business_Demography_outputs_using_the_Noise_for_Counts_and_Magnitudes_NCM_me

thod?_iepl%5BviewId%5D=GwgMrIxMHU8nyj0nX660tuHJ&_iepl%5BprofilePublicationItemVariant%5D=default&_iepl%5Bcontexts%5D%5B0%5D=prfpi&_iepl%5BtargetEntityId%5D=PB%3A311734931&_iepl%5BinteractionType%5D=publicationTitle

Morgan, R. (2016). *Geospatial Confidentiality*. Internal Stats NZ presentation.

NZ Government. (2013). *Statistics Act 1975*. http://www.legislation.govt.nz/act/public/1975/0001/latest/DLM430705.html.

Stats NZ. (2017). *Integrated Data Infrastructure.* http://www.stats.govt.nz/browse_for_stats/snapshots-of-nz/integrated-data-infrastructure

Stats NZ. (2016). *Microdata Output Guide*. http://www.stats.govt.nz/tools_and_services/microdata-access/data-lab/microdata-output-guide

Wallgren, A. & Wallgren, B. (2014). *Register-based Statistics; statistical methods for administrative data.* 2nd edition, Wiley series in survey Methodology