Establishing an Automated Confidentiality Service in Stats NZ

Antony Gomez, Frances Krsinich and Allyson Seyb

Introduction

- Stats NZ has a vision of unleashing the power of data to change lives.
- Increase the value of data and move towards a more open data environment.
- 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.
- Bound by the Statistics Act 1975 to maintain privacy and confidentiality by not disclosing information about an individual or business.
- To move to a more open and accessible data environment there is a

What is an Automated Confidentiality Service?

- It is about how to deliver confidentialised output / tables.
- It is not the underlying methodology for confidentialising.
- It is a service and not a tool. ACS can have many tools.
- It requires assigning a random number [0,1] to each unit record (person, business or location) to achieve consistency.
- Reduce manual application and checking of confidentiality rules.
- Significant time savings. Quicker turnaround e.g. Customer services – initial request to output table.

What is an Automated Confidentiality Service?

- More consistent and accurately applied confidentiality procedures.
- More efficient and accurate analysis.
- More open and accessible data – **Unleashing the power of data**
  - Increasing the value of data.
  - Increase the quality of research.
- Greater use by the public (majority compared to researchers) through web-based query and visualisation tools.

What is an Automated Confidentiality Service?

- Reduces barrier to entry. (Don’t need to understand confidentiality to get safe outputs).
- Increases data usage.
- Incentive for people to share their data with us.
- Display leadership / expertise in providing a Confidentiality service for other organisations.
- Fewer resources needed to train researchers and check results.

Noise for Counts and Magnitudes (NCM) method

- Requires a random seed at the unit record (individual, business, address) level. Uniformly distributed on [0,1].
- Confidentialised cell counts are based on a cell-level random number
- Confidentialised magnitudes are derived using a noise multiplier (US Census ‘EZS’ method).

Noise for Counts and Magnitudes (NCM) method – Counts

- For table cell counts, the cell-level random seed is derived from summing the random seeds of the contributing unit records for that cell.
- Mod1 of sum of random seeds or drop integer part. Also uniformly distributed on [0,1].
- Random Rounding to base 3 (RR3) is based on cell-level seed - ‘fixed’ RR3 (FRR3)
  - If cell_ran <= 2/3 then round to nearest multiple of 3.
  - If cell_ran > 2/3 then round to the next nearest multiple of 3.

Noise for Counts and Magnitudes (NCM) method – Counts example

Stats NZ
Tatauranga Aotearoa

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**

Advantage of Fixed Random Rounding (FRR3)

- Simple, easy to apply.
- Cell with same contributors always rounded consistently
- Consistent for the same cell in different structured table.
- Familiar to users.
- Non-additivity means cell-level suppression doesn’t require secondary suppression.

- Perturbation applied on the outputs.

Noise for Counts and Magnitudes (NCM) method – Magnitudes

- Noise multiplier is applied to the magnitudes at the unit record level.
- For a minimum of 10% perturbation use the random seed to determine level of noise within 10% for that unit record.
  - If ran < 0.5 then noise_mult = 0.9 – (0.5 – ran)/100.
  - If ran >= 0.5 then noise_mult = 1.1 + (ran – 0.5)/100.
  - Multiply magnitude by noise_mult
- Magnitude cell values are derived by summing the perturbed magnitudes of the contributing unit records.

Noise for Counts and Magnitudes (NCM) method – Magnitudes example

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

Noise for Counts and Magnitudes (NCM) method – Magnitudes example

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**

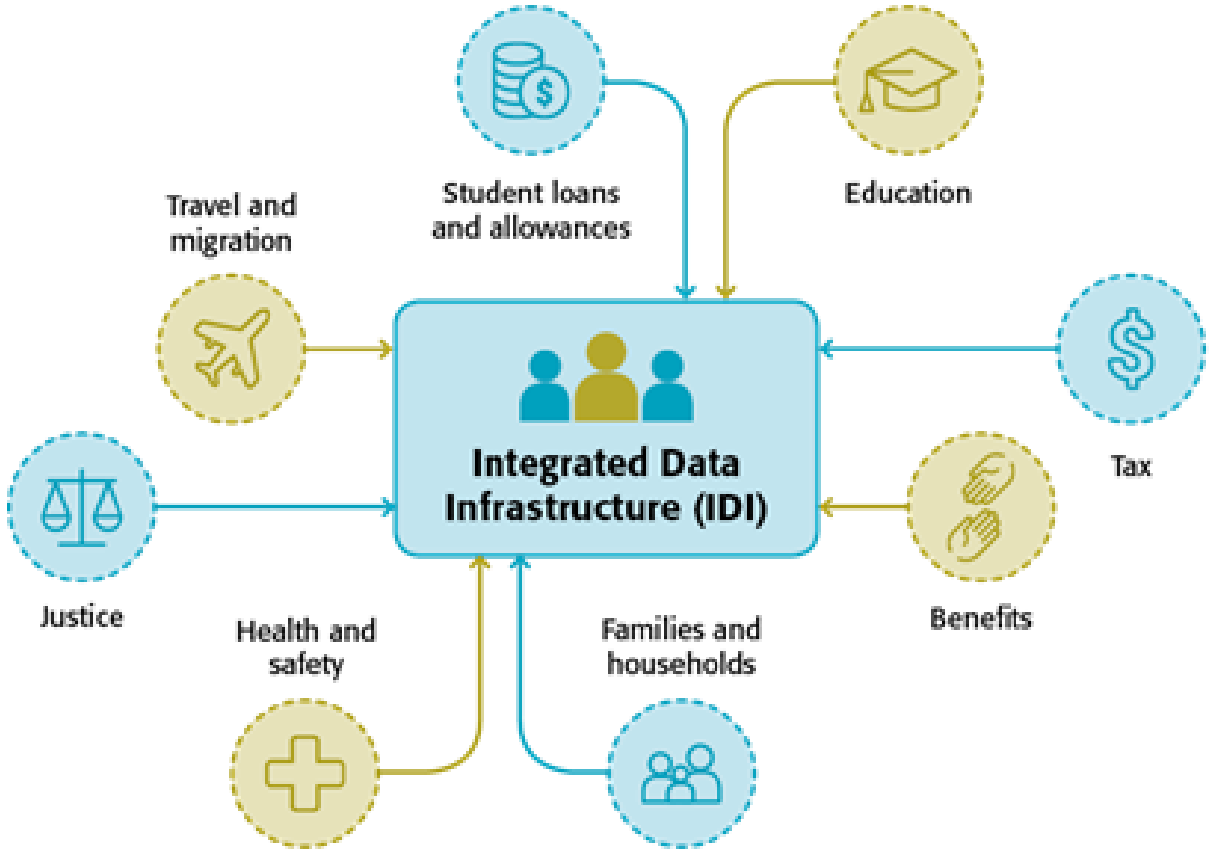
Percentage difference between original and noised employee counts

	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

Advantages of Noise for Magnitudes

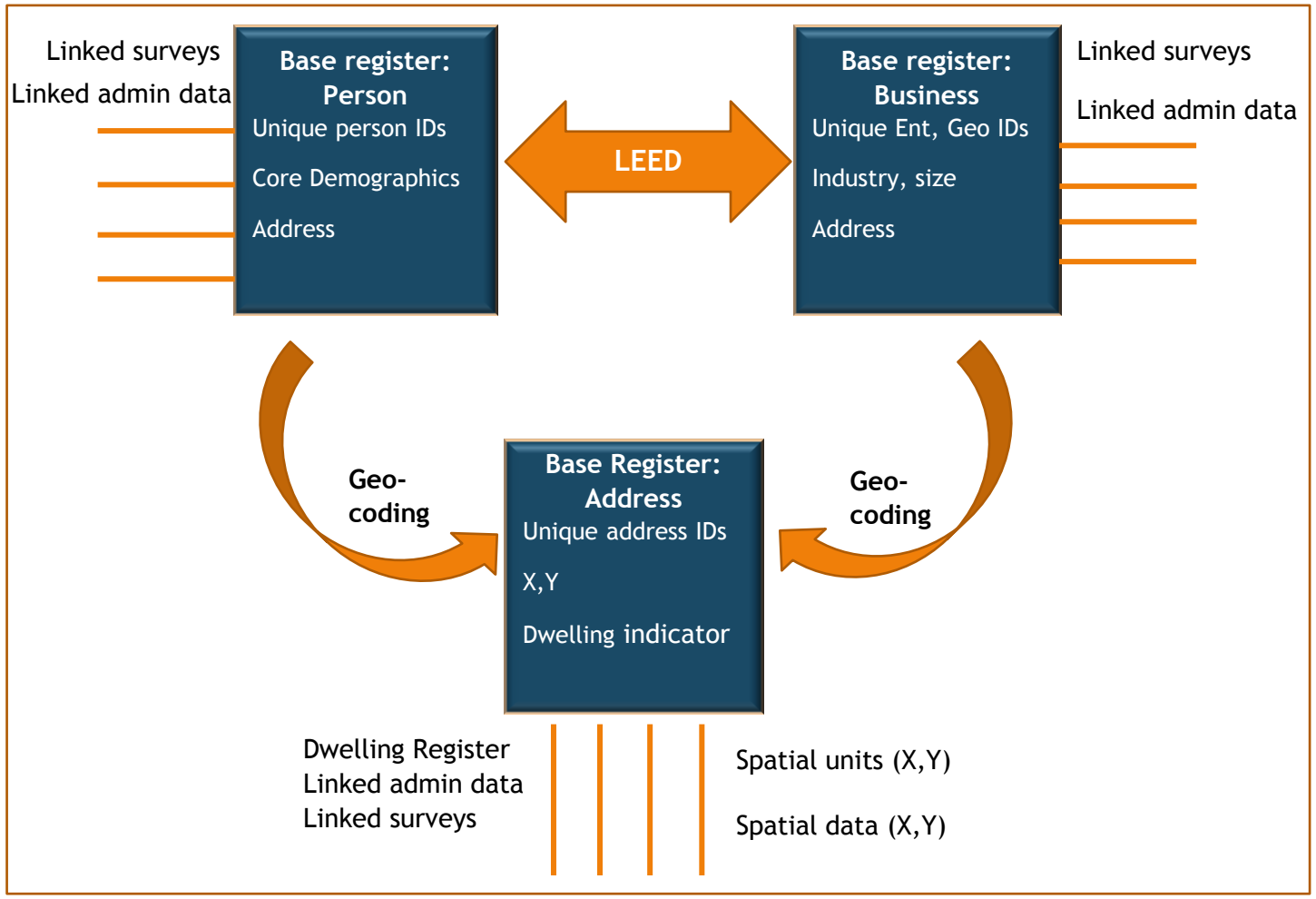
- Simple, easy to apply.
- Same cell always noised the same way.
- Noise targeted towards sensitive cells.
- Noise is averaged out for cells with a large number of contributing unit records.
- Preserves additivity.
- No suppression required (in particular, no sacrificing non-sensitive cells to secondary suppression).
- Perturbation is applied on the input.

Integrated Data Infrastructure (IDI)



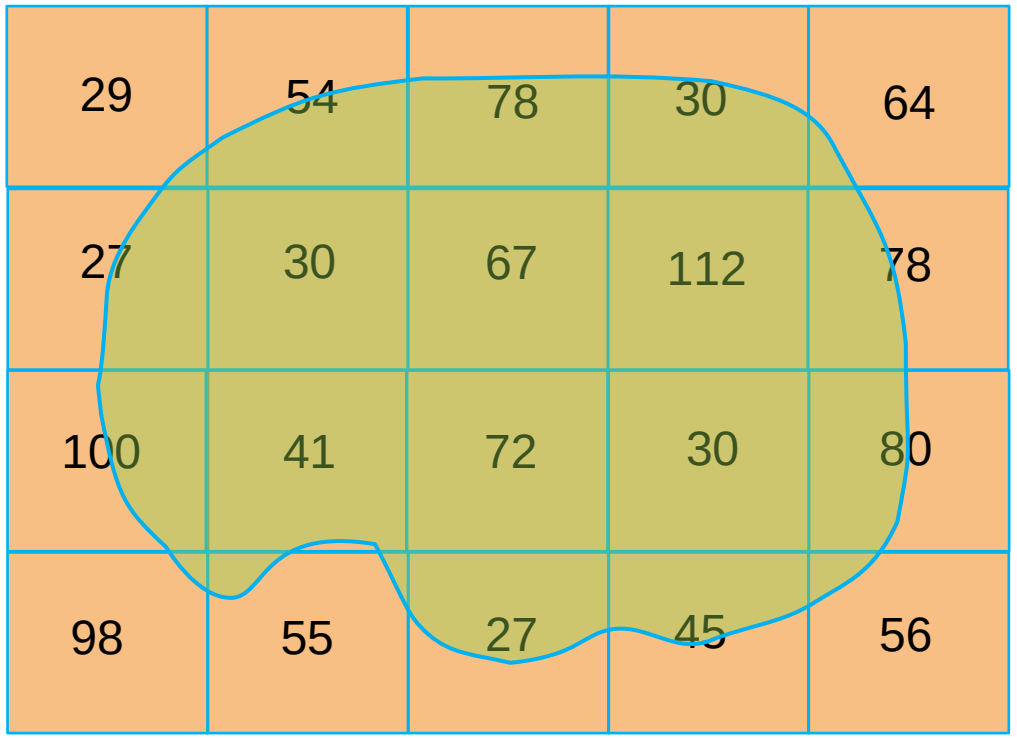
- Large secure database of linked microdata used for social research.
- Stats NZ survey data, including Census, government agencies' and some non-government organisations administrative data.
- Approved researchers in a secure Datalab environment.
- Checked by Stats NZ staff – Phase 1 and Phase 2 checks.
- Automating the checking process.

Registered-based Statistical System (RSS)



- Three base registers – Person, Business, and Address.
- Random number assigned to each statistical unit in each base registers.
- Random number is carried through to the linked data.

Geospatial confidentiality – customised geographies



Key

Meshblock

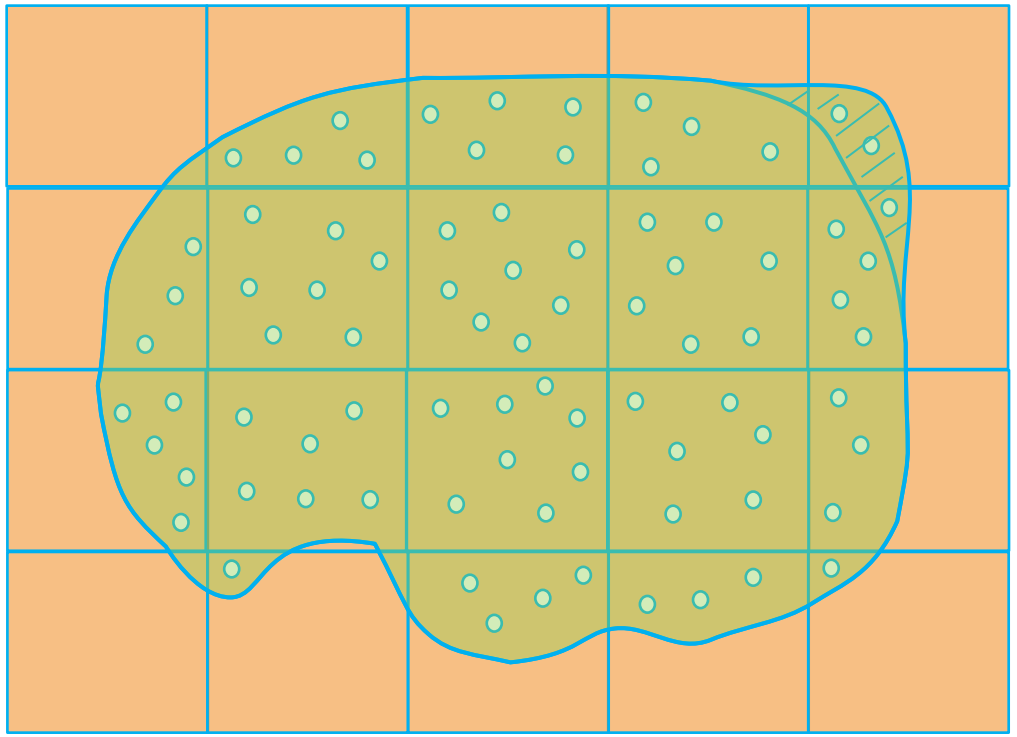
X

UR Population

Area of interest

- Customised output geographies – non-meshblock defined areas.
- Meshblock 'size'
  - Should be < 80 dwellings.
  - Split to 30 – 60 dwellings.
- Meshblock is the lowest geographical output level, rounded to base 3.

Geospatial confidentiality – customised geographies

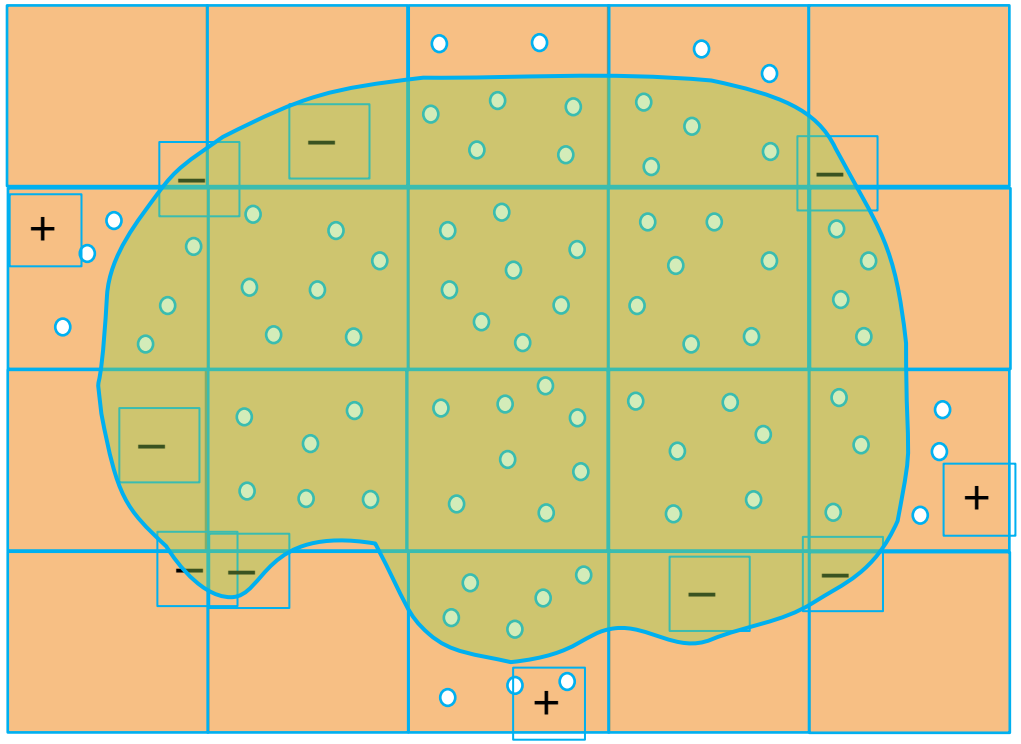


Key

- Meshblock boundary
- Area of interest
- Dwelling
- ‘difference’

- Changing the boundary slightly, leads to differencing issues.
- How to ensure Confidentiality?

Geospatial confidentiality – customised geographies



Key:

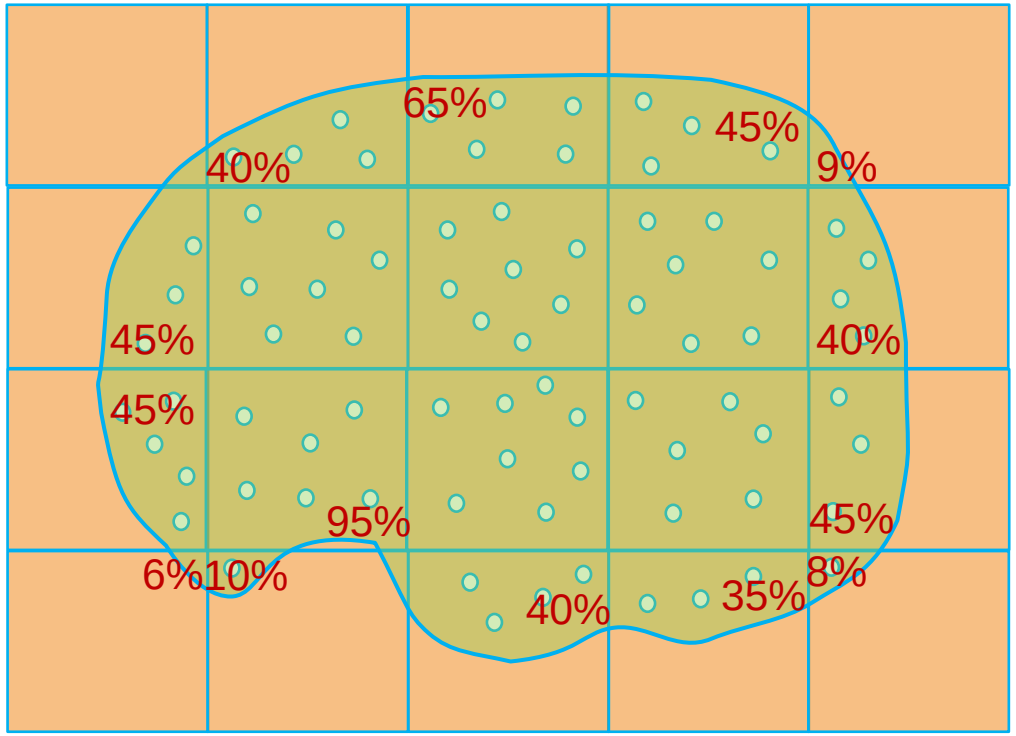
Meshblock boundary

Area of interest

Dwelling

- Reassign dwellings from one meshblock to another
- Equivalent to realigning to the meshblock boundaries.
- Difficult to manage or automate.

Geospatial confidentiality – customised geographies



Key:

- Meshblock boundary
- Area of interest
- Dwelling
- Y% Proportion of dwellings

- Dwellings are assigned a fixed random number
- For each meshblock determine the proportion of dwellings that lie inside area of interest.

Geospatial confidentiality – customised geographies

- Select all the dwellings whose random number is less than or equal to the proportion of dwellings in the meshblock.
  - E.g. if 75% of dwellings lie within the area of interest then select all dwellings with random numbers <= 0.75.
- Randomly selecting dwellings within the meshblock to represent the proportion - estimation.
- Selected dwelling could lie outside the area of interest.
- Changing the boundary changes the proportion and selection
- Differencing no longer an issue.
- Confidentiality is preserved and output remains consistent.

Random Numbers

- But the random number is assigned to confidentialise the number of dwellings not for selection of dwellings.
- Using the same random number for both could lead to some form of bias.
- We know the digits of a random number are themselves random.
- Multiply the random number by 10 then mod1 to drop the integer to create a new random number for the selection.
- Use the fixed random number for other purposes.

Geospatial confidentiality – customised geographies

- To count (estimate) the number of persons within a non-meshblock area, select the dwellings representing the area.
- Count the number of persons in each of those dwellings.
- Use the fixed random number for those persons to determine the cell values then use FRR3 to round to base 3.

Questions?

Thank you.