Statistical Disclosure Control on 2021 EU Census grid data – the fast track

2017 EFGS conference, Dublin

Ekkehard PETRI, Fabian BACH Eurostat

(Ekkehard.Petri@ec.europa.eu)

Overview

- Planned EU regulation for 2021 population grids
- User requirements for Statistical Disclosure Control (SDC) on population grids
- Overview of possible SDC methods
- A very fast tour of the cell key method

2021 Population census – major SDC challenge

- (planned) **1 km² grids** with 13 key variables per grid square

➔Many grid cells with few persons

Eurostat Grant "Harmonised protection of 2021 Census data"

Topic	Breakdown categories CIR-1	Breakdown categories CIR-4	Description	STAT.G.
GEO.		GEO.G.x. GEO.G.y.	(See section **Error! Reference source not found.**)	
SEX.	SEX.0. SEX.1. SEX.2.		Total population Male Female	0. 1. 2.
AGE.		AGE.G.1. AGE.G.2. AGE.G.3.	Under 15 years: equal to AGE.L.1. in CIR-1 15 to 64 years: sum of AGE.L.2.-4. in CIR-1 65 years and over: sum of AGE.L.5.-6. in CIR-1	3. 4. 5.
CAS.	CAS.L.1.1.		Employed persons (see details in section **Error! Reference source not found.**)	6.
POB.	POB.L.1. POB.L.2.1. POB.L.2.2.		Place of birth in reporting country Place of birth in other EU Member State Place of birth elsewhere	7. 8. 9.
ROY.	ROY.1. ROY.2.1. ROY.2.2.		Usual residence unchanged Move within the reporting country Move from outside the reporting country	10. 11. 12.

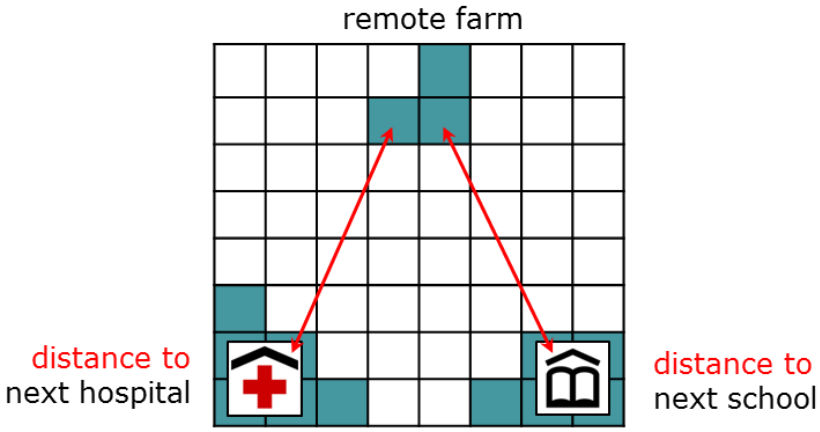
Specific requirements for SDC on grid statistics

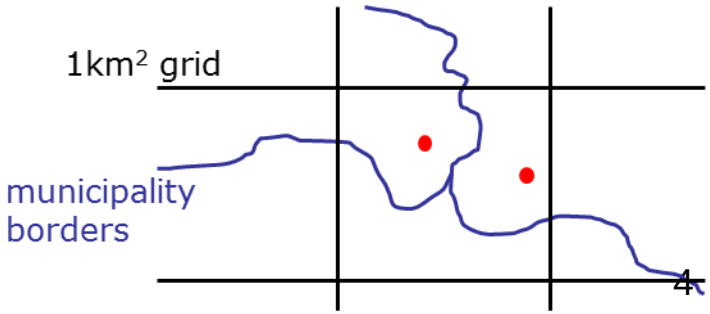
Working Assumption: Main application of grid statistics are accessibility studies for planning and analysis

⇒ need to know where people live and if few or many

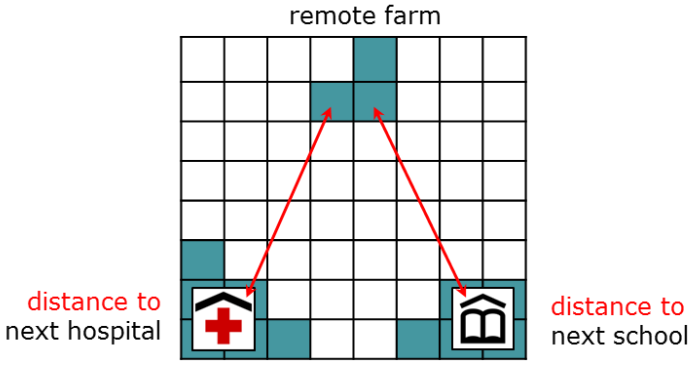
⇒ Exact count (e.g. 4 or 5) not essential

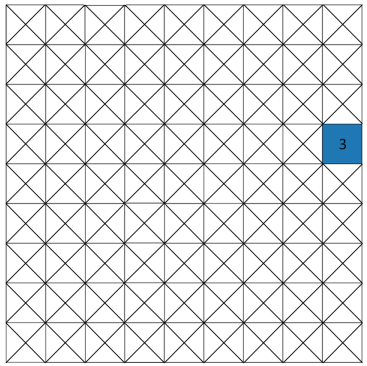
- SDC should minimise information loss, in particular as regards inhabited and non-inhabited
- Recommended method should be applicable in all NSIs possibly with different parameters

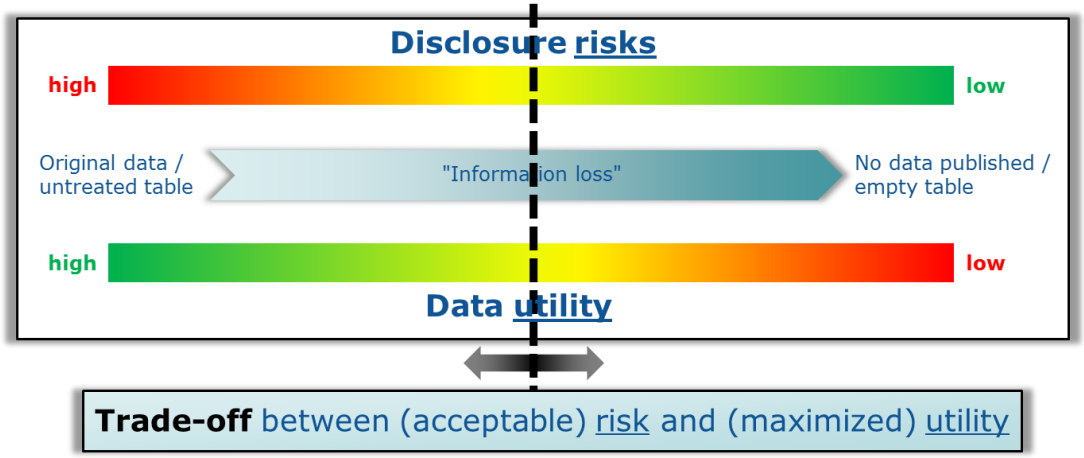




The eternal SDC conflict

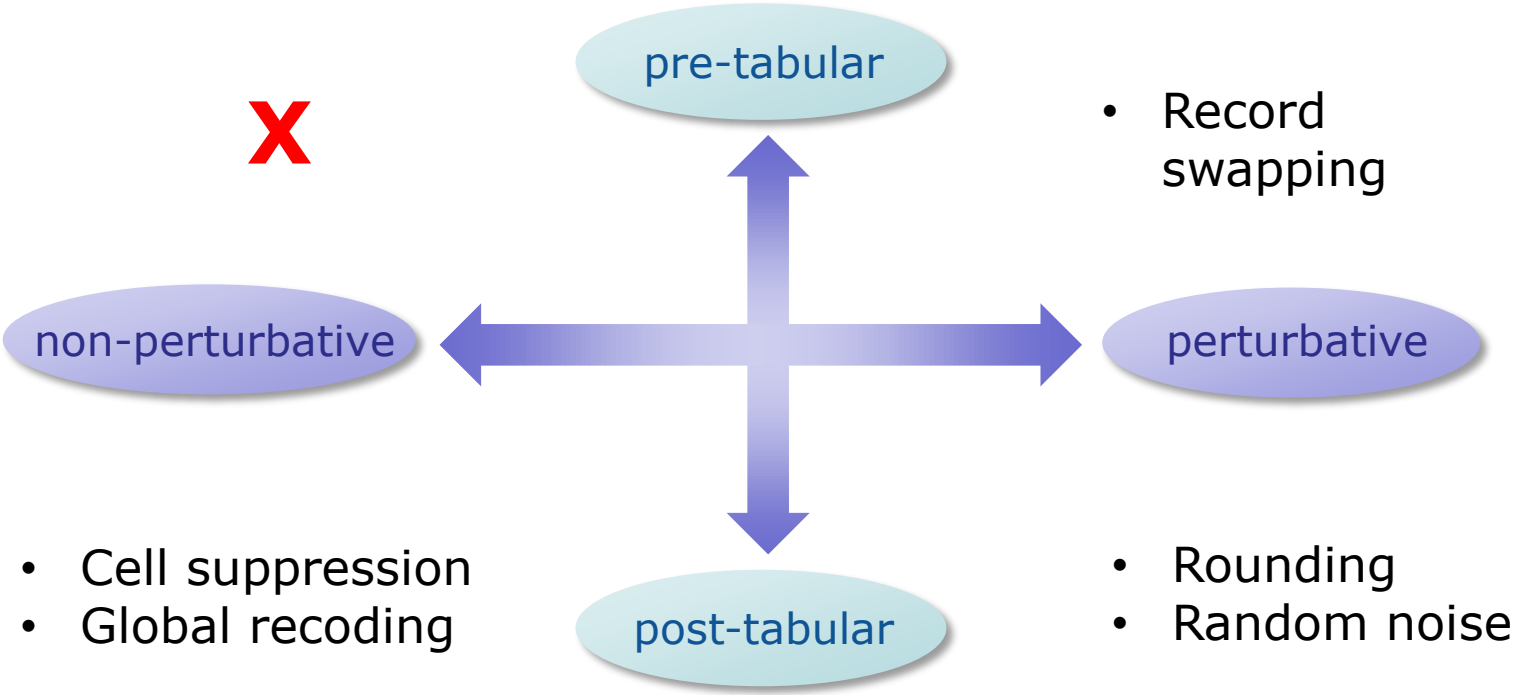






Confidentiality methods considered

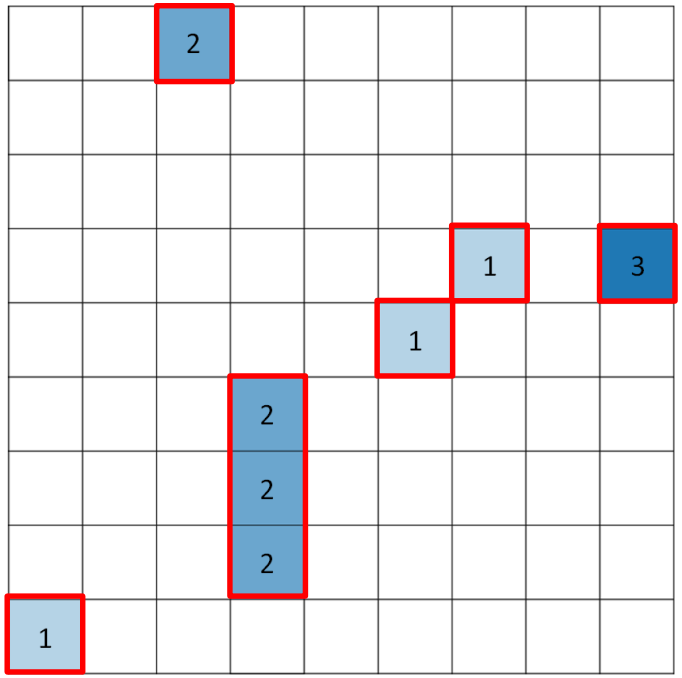
- Confidentiality methods can be categorised:



Effects of confidentiality treatment

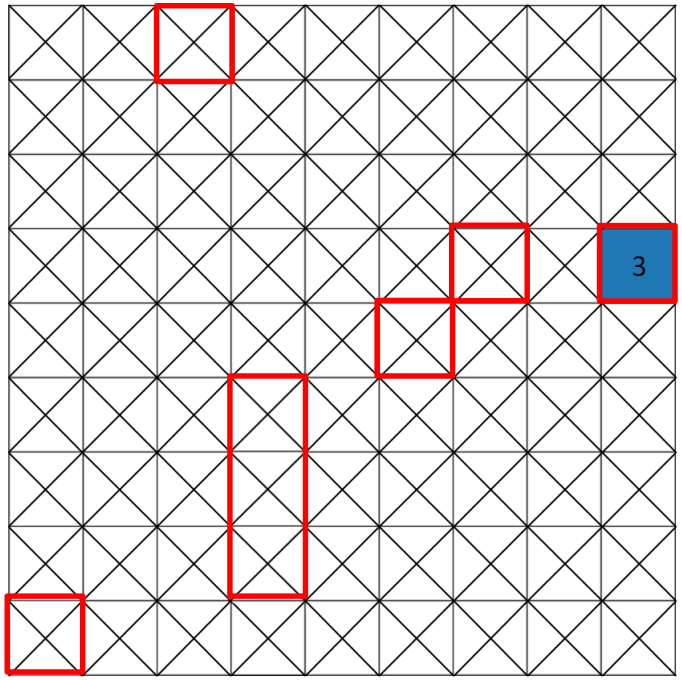
Consider sparsely populated region:

populated squares:
8



Effects of confidentiality treatment

populated squares:
8 → ???



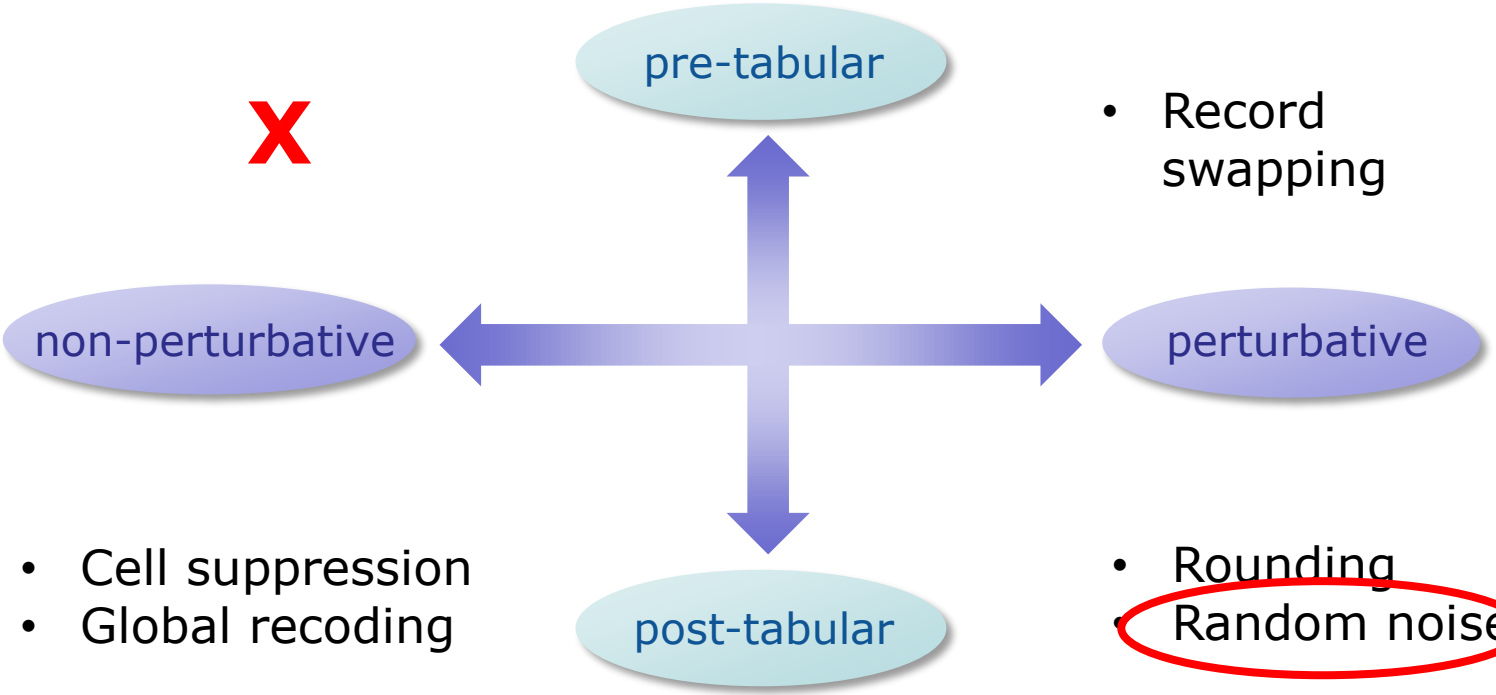
Cell suppression (all counts <3)

Requirements for the *2021 EU Census (including grids)*

- Keep regulated table structure (→ no recoding)
- Minimal information loss, in particular on inhabited or not (→ no suppression)
- Control geographic differencing (→ record swapping insufficient)
- Minimize inconsistencies (→ no suppression, no recoding)

Confidentiality methods considered

- Confidentiality methods can be categorized:



Random noise – the cell key method

- Variant of **random noise** developed by the *Australian Bureau of Statistics:* **strictly consistent** through *cell keys*, e.g. 1…200

Ptable¤		Cell·key··(1-200)¤							¤
		1¤	2¤	3¤	…¤	62¤	…¤	200¤	¤
Cell· Value¤	1¤	¤	+1¤	¤	¤	-1¤	¤	¤	¤
	2¤	-1¤	¤	¤	¤	¤	¤	¤	¤
	3¤	¤	¤	¤	¤	¤	¤	-1¤	¤
	4¤	+1¤	¤	¤	¤	-1¤	¤	¤	¤
	5¤	¤	¤	-1¤	¤	¤	¤	¤	¤
	…¤	¤	¤	¤	¤	¤	¤	¤	¤

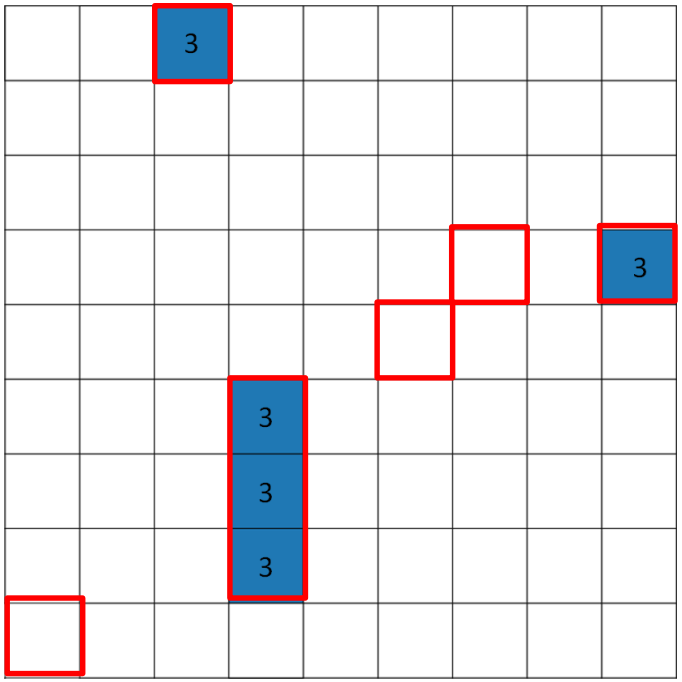
Age \ Sex	M	F
0-15	.	.
16-24	.	1
25-34	.	.
…	…	…

-1

Age \ Sex	M	F
0-15	.	.
16-24	.	0
25-34	.	.
…	…	…

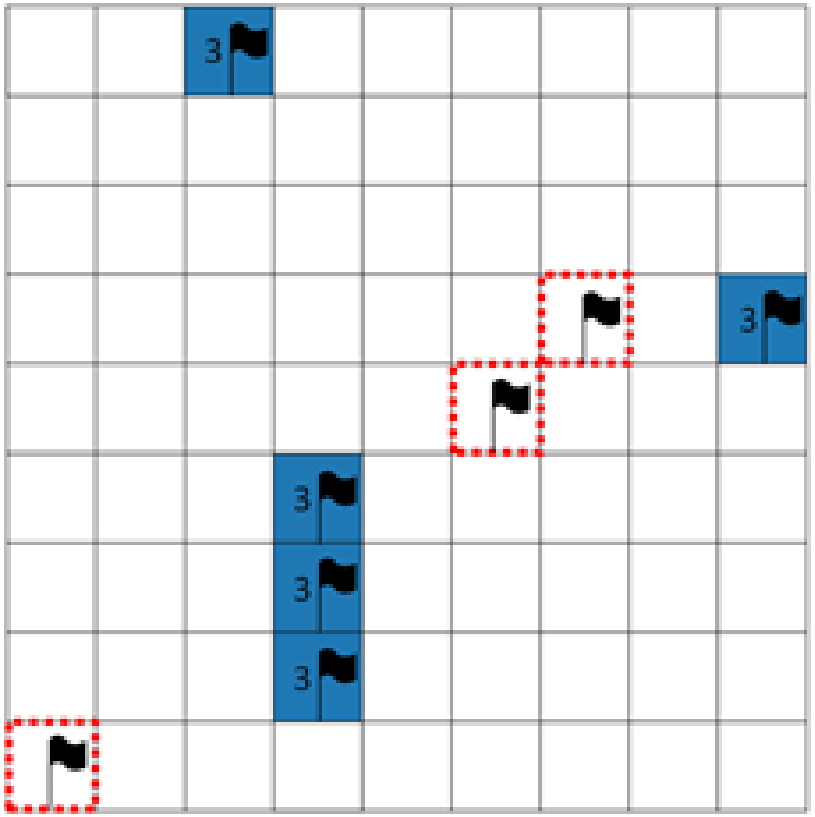
Effects of confidentiality treatment

populated squares:
8 → 5



Small perturbation (rounding, random noise)

Draft compromise for 2021



Sex / Age	M	F
0-15	.	.
16-24	.	Inh.
25-34	.	.
...	...	...

Conclusion – SDC for 2021 population grids

- Consistency between hyper-cubes and grid data
- Possible to combine record swapping and cell key method
- Possible for all Member States - specific parameters (p-table design)
- Compromise between disclosure control and information loss thanks to inhabited flag
- All the details in the Annex…


European
Commission

Annex – detailed example of the cell key method

- Variant of **random noise** developed by the *Australian Bureau of Statistics:* **strictly consistent** through *cell keys*, e.g. 1…200

Cell key method

- Variant of **random noise** developed by the *Australian Bureau of Statistics:* **strictly consistent** through *cell keys*, e.g. 1…200
- **Algorithm** – step 1: assign each record a random number (record key, "Rkey") between 1 and max. cell key (here 200)

ID	Region	Sex	Age	…
1	A	M	31	…
2	A	F	47	…
3	B	M	22	…
…	…	…	…	…

Cell key method

- Variant of **random noise** developed by the *Australian Bureau of Statistics:* **strictly consistent** through *cell keys*, e.g. 1…200
- **Algorithm** – step 1: assign each record a random number (record key, "Rkey") between 1 and max. cell key (here 200)

ID	Region	Sex	Age	…	*Rkey*
1	A	M	31	…	54
2	A	F	47	…	104
3	B	M	22	…	93
…	…	…	…	…	…

Cell key method

- Variant of **random noise** developed by the *Australian Bureau of Statistics:* **strictly consistent** through *cell keys*, e.g. 1…200
- **Algorithm** – step 2: Create the tables. For each cell, sum all *Rkeys* modulo 200

Age \ Sex	M	F
0-15	.	.
16-24	.	4
25-34	.	.
…	…	…

Cell key method

- Variant of **random noise** developed by the *Australian Bureau of Statistics:* **strictly consistent** through *cell keys*, e.g. 1…200
- **Algorithm** – step 2: Create the tables. For each cell, sum all *Rkeys* modulo 200

Age \ Sex	M	F
0-15	.	.
16-24	.	4
25-34	.	.
…	…	…

ID	*Rkey*
2	104
4	61
56	7
72	90

Cell key method

- Variant of **random noise** developed by the *Australian Bureau of Statistics:* **strictly consistent** through *cell keys*, e.g. 1…200
- **Algorithm** – step 2: Create the tables. For each cell, sum all *Rkeys* modulo 200

Age \ Sex	M	F
0-15	.	.
16-24	.	**4**
25-34	.	.
…	…	…

ID	*Rkey*
2	104
4	61
56	7
72	90

Ckey = Σ(*Rkey*) mod 200
= 262 mod 200
= **62**

Cell key method

- Variant of **random noise** developed by the *Australian Bureau of Statistics:* **strictly consistent** through *cell keys*, e.g. 1…200
- **Algorithm** – step 3: Use pre-defined perturbation table ("ptable") to get noise value

Ptable		Cell key (1-200)						
		1	2	3	…	62	…	200
Cell Value	1		+1					
	2	-1						
	3							-1
	4	+1				+1		
	5			-1				
	…							

Cell key method

- Variant of **random noise** developed by the *Australian Bureau of Statistics:* **strictly consistent** through *cell keys*, e.g. 1…200
- **Algorithm** – step 4: Add noise value to cell

Age \ Sex	M	F
0-15	.	.
16-24	.	4
25-34	.	.
…	…	…

+1

Cell key method

- Variant of **random noise** developed by the *Australian Bureau of Statistics:* **strictly consistent** through *cell keys*, e.g. 1…200
- **Algorithm** – step 4: Add noise value to cell

Age \ Sex	M	F
0-15	.	.
16-24	.	4
25-34	.	.
…	…	…

+1

Age \ Sex	M	F
0-15	.	.
16-24	.	5
25-34	.	.
…	…	…

Cell key method

- Variant of **random noise** developed by the *Australian Bureau of Statistics:* **strictly consistent** through *cell keys*, e.g. 1…200
- **Algorithm** – step 4: Add noise value to cell

Age \ Sex	M	F
0-15	.	.
16-24	.	4
25-34	.	.
…	…	…

+1

Age \ Sex	M	F
0-15	.	.
16-24	.	5
25-34	.	.
…	…	…

- Pseudo-random: will **always** be **+1 for this cell**, due to ***Ckey***
- **Fixed noise variance** through ptable design