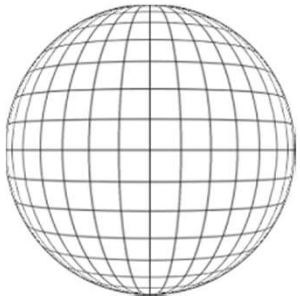
Discrete Global Grid System (DGGS) and Confidentiality

Standards: Bringing geospatial information and statistics together



Ana Santos | Statistics Portugal
Vilni Verner Holst Bloch | Statistics Norway

Instituto Nacional de Estatística
Statistics Portugal

Statistisk sentralbyrå
Statistics Norway

Content

- Global Forum for Geography and Statistics
- DGGS and standardised grids
- Grids and confidentiality

Global Forum for Geography and Statistics

A voluntary task force connected to EFGS since 2016

GFGS@ EUROPEAN FORUM FOR GEOGRAPHY AND STATISTICS

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

eurostat

Statistisk sentralbyrå
Statistics Norway

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

SOUTHERN TERRA COGNITA LABORATORY

JRC
EUROPEAN COMMISSION

CIESIN
Columbia University

KRIHS
Korea Research Institute for Human Settlements

Stats NZ
Tatauranga Aotearoa

United States Census Bureau

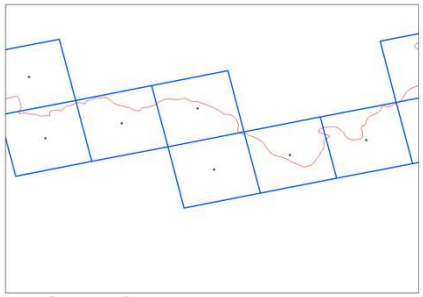
IBGE
Instituto Brasileiro de Geografia e Estatística

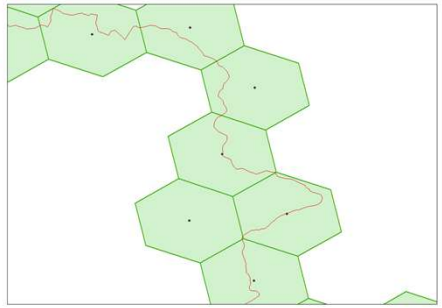
Office for National Statistics

INE Instituto Nacional de Estadística

Australian Bureau of Statistics

Global Forum for Geography and Statistics



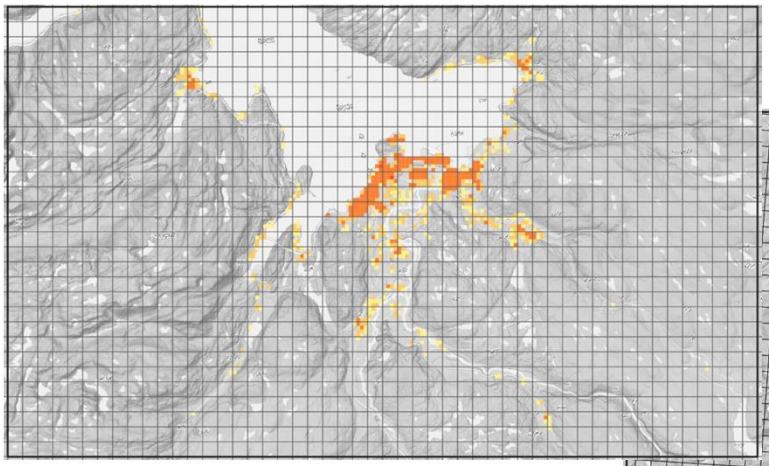


- **The 2018 grid challenge:**
  - Make population on hexgrid
  - Results after 2-3 weeks: 15 countries and territories
- **The 2019 border soup experiment:**
  - Compare square and hex grids
  - Hexagons slightly more efficient in sampling and presenting point data

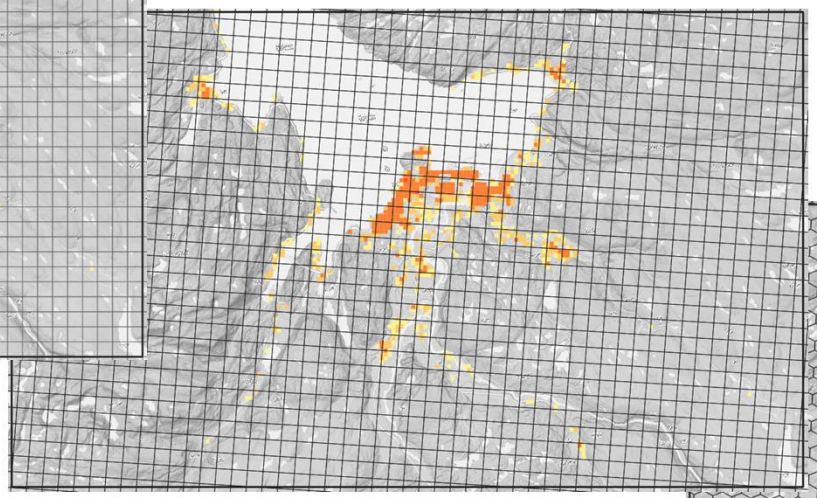
- **2021 Grids and confidentiality:**
  - Challenges and methods

DGGS and standardised grids

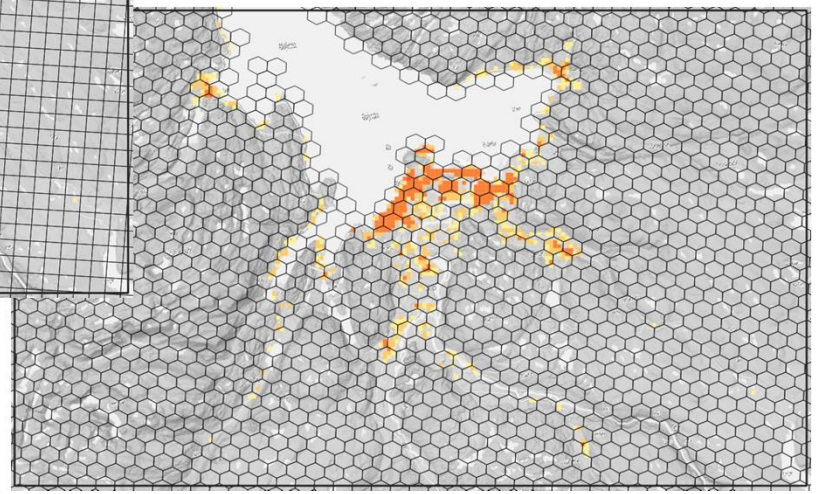
National grid



European census grid



Global DGGS grid



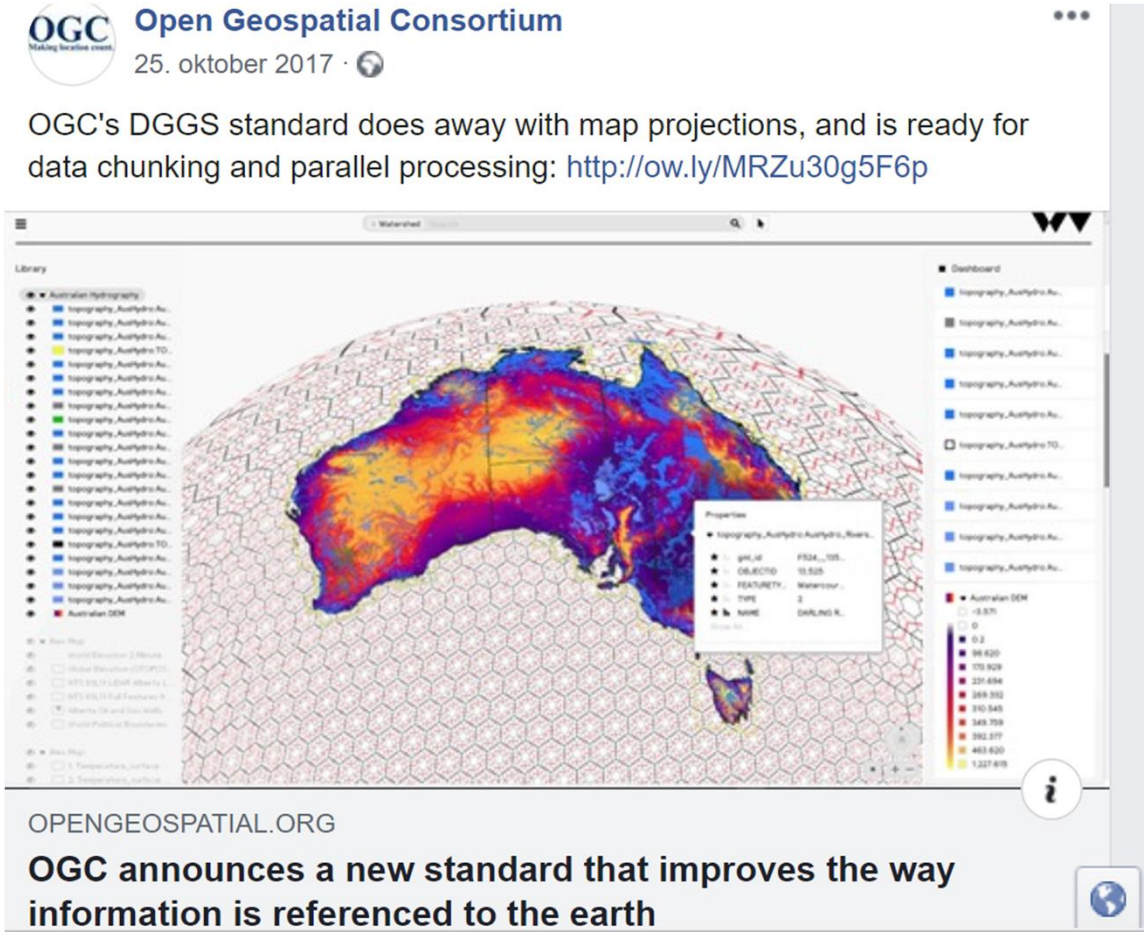
1 – 4 Inhabitants
5 – 19 Inhabitants
20 – 199 Inhabitants
200 – 499 Inhabitants
500 – 5 000 Inhabitants
> 5000 Inhabitants

Census 2021 (INSPIRE grid):
Total population
Sex
Age groups (0-14, 15, 64, 65+)
Birth country (Norway/EU/other)
Living place 1 year before census
(unchanged, internally migrated, migrated from abroad)

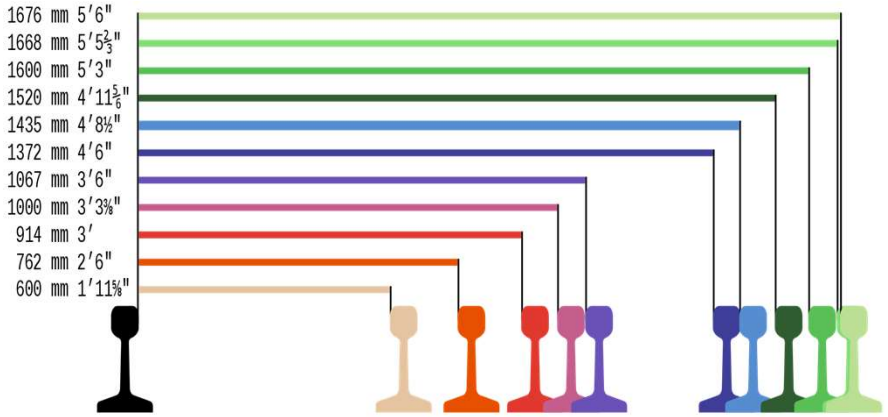
Instituto Nacional de Estatística
Statistics Portugal

Statistisk sentralbyrå
Statistics Norway

Discrete Global Grid System



The railroad standard for grids



Discrete Global Grid System

The Minecraft rules

How to build the world

R 1 Reference frame and algorithms
R 2 Domain completeness
R 3 Position uniqueness
R 4 Resolutions
R 5 Successive completeness
R 6 Cell shape
R 7 Cell area
R 8 Equal area
R 9 Initial tessellation
R 10 Refinement method
R 11 Spatial reference
R 12 Unique addresses
R 13 Cell location
R 14 Quantization methods

The Pokémon Go of grids

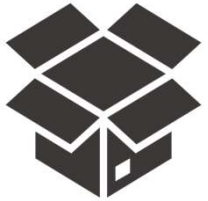
How to move around

R 15 Navigation operations
R 16 Spatial analysis

The Black Box rules

How to ask

R 17 Data query
R 18 Data transfer



Mix of grid and confidentiality

- Confidentiality not restricted to grids
- Could be any mix statistics by regional distribution
- But mix of statistics on geographical grids are particularly challenging
- Start with example with new disclosure rules in Norway
- Give some examples on methods for handling statistical and geospatial information

Example: disclosure rules in Norway

- No exact figures for 1-9 persons in a grid cell
- Applies to any size of grid cell
- Consequences for gridded population data sets?

Example: disclosure rules in Norway

2020				
Supression 1. step	Populated Grid cells (N)	Grid cells Unsuppressed (N)	Grid cells Suppressed (N)	Grid cells Suppressed (%)
ID0250M	222 247	69 842	152 405	**68,6**
ID1000M	54 967	31 334	23 633	43,0
ID5000M	7 621	6 177	1 444	18,9

2. step

Additional 19 373 (**8,7 %**) 250m grid cells must be suppressed because only

one 250m grid cell is suppresed within a 1km grid cell.

3. step

Further suppression in 5 km grid cells …

Example: disclosure rules in Norway

	Grid cells (N)			Persons (N)		
pop_tot	SSB0250M	SSB1000M	SSB5000M	SSB0250M	SSB1000M	SSB5000M
In all (1-9)	152 405	23 633	1 444	577 589	99 769	6 071
1	25 459	3 770	274	25 459	3 770	274
2	32 722	4 127	225	65 444	8 254	450
3	21 022	2 923	165	63 066	8 769	495
4	21 103	2 842	169	84 412	11 368	676
5	16 220	2 462	159	81 100	12 310	795
6	12 516	2 204	114	75 096	13 224	684
7	9 844	1 971	118	68 908	13 797	826
8	7 567	1 729	109	60 536	13 832	872
9	5 952	1 605	111	53 568	14 445	999

About 600 000 residents must be suppressed at country level.
At regional level the share of suppressed 250 m grid cells will be very high.

Example: disclosure rules in Norway

- In practise not possible to publish on other grid cell sizes
  (given simple suppression)
- Loss of information
- How to handle without totally damaging usefulness/completeness of dataset?

Challenge and Solution

- Take care of privacy without damaging data too much
- Principles for masking
- Methods for masking
- Different requirements to data sets

Principles for masking

1. Masking must not be reversible.	However you mask your data, it should never be possible to use it to retrieve the original sensitive data.
2. Results must be representative of the source data.	The reason to mask data instead of just generating random data is to provide non-sensitive information that still resembles production data for development and testing purposes. This could include geographic distributions, credit card distributions (perhaps leaving the first 4 numbers unchanged, but scrambling the rest), or maintaining human readability of (fake) names and addresses.
3. Referential integrity must be maintained.	Masking solutions must not disrupt referential integrity — if a credit card number is a primary key, and scrambled as part of masking, then all instances of that number linked through key pairs must be scrambled identically.
4. Only mask non-sensitive data if it can be used to recreate sensitive data.	It isn't necessary to mask everything in your database, just those parts that you deem sensitive. But some non-sensitive data can be used to either recreate or tie back to sensitive data. For example, if you scramble a medical ID but the treatment codes for a record could only map back to one record, you also need to scramble those codes. This attack is called inference analysis, and your masking solution should protect against it.
5. Masking must be a repeatable process.	One-off masking is not only nearly impossible to maintain, but it's fairly ineffective. Development and test data need to represent constantly changing production data as closely as possible. Analytical data may need to be generated daily or even hourly. If masking isn't an automated process it's inefficient, expensive, and ineffective.

Securosis (). Understanding and Selecting Data Masking Solutions

Methods for masking

Substitution	Substitution is simply replacing one value with another.
Redaction/Nulling	This substitution simply replaces sensitive data with a generic value, such as ‘X’.
Shuffling	Shuffling is a common randomization technique for disassociating sensitive data relationships (e.g., Bob makes $X per year) while preserving aggregate values.
Blurring	Taking an existing value and alter it so that the value falls randomly within a defined range.
Averaging	Averaging is an obfuscation technique where individual numbers are replaced by a random value, but across the entire field, the average of these values remains consistent.
De-identification	A generic term for to any process that strips identifying information, such as who produced the data set, or personal identities within the data set.
Tokenization	Tokenization is substitution of data elements with random placeholder values. Tokens are non-reversible because the token bears no logical relationship to the original value.
Format Preserving Encryption	Encryption is the process of transforming data into an unreadable state. Unlike the other methods listed, the original value can be determined from the encrypted value, but can only be reversed with special knowledge (the key).

Limitations for masking

Format Preservation: The mask must produce data with the same structure as the original data. E.g. if the original data is 2-30 characters long, the mask should produce data 2-30 characters long.

Data Type Preservation: With relational data storage it is essential to maintain data types when masking data from one database to another. Relational databases require formal definition of data columns and do not tolerate text in number or date fields.

Gender preservation: When substituting names, male names are only substituted with other male names, and similarly female with only female names.
Semantic Integrity: Databases often place additional constraints on data they contain such as a LUHN check for credit card numbers, or a maximum value on employee salaries. In this way you ensure both the format and data type integrity, but the stored value makes sense in a business context as well.
Referential Integrity: Shuffling or substituting key data values can destroy these references (relationships). Masking technologies must maintain referential integrity when data is moved between relational databases, or cascading values from one table to another. This ensures that loading the new data works without errors and avoids breaking applications which rely on these relationships.
Aggregate Value: The total and average values of a masked column of data should be retained, either closely or precisely.
Frequency Distribution: In some cases users require random frequency distribution, while in others logical groupings of values must be maintained or the masked data is not usable. The ability to mask data while maintaining certain types of patterns is critical for maintaining the value of masked data for analytics.
Uniqueness: Masked values must be unique. As an example, duplicate SSNs are not allowed when uniqueness is a required integrity constraint. This is critical for referential integrity, as the columns used to link tables must contain unique values.

2020-NO-CENSUS-GEO

- EU funded grant at Statistics Norway for confidentiality on grid data
- Goal: compare masking methods (cell-key and small count rounding) for grid census data, provide recommendations for best-practice
- based on R packages SSBcellKey[1] and SmallCountRounding[2]

[1]https://github.com/statisticsnorway/SSBcellKey
[2]https://cran.r-project.org/web/packages/SmallCountRounding/index.html

Method examples

- **Double Pseudonomized Rank Shuffling (DPRS)**
- **Target record swapping**
- **Cell key method**

Double Pseudonomized Rank Shuffling (DPRS)

- Mask all cells with less than X persons and scramble at smallest grid cell level in one grid system
- Use double pseudonymised key (PNR1 and PNR)
- Make new key by rank ordering (PNR1->RANK1)
- Make new key by rank ordering (PNR2->RANK2)
- Shuffle information using RANK1 and RANK2
- Aggregate to wanted grid system

DPRS illustrated

PSEUDONOMIZED PID NR 1

PSEUDONOMIZED PID NR 2

PID	PNR1	PNR2	VAR1	VAR2	VAR3	ADR	GRID-ID	RANK1	RANK2
A			1	M	0	A	A		
B			32	M	2	A	A		
C			23	K	1	B	A	13 224	44 278
D			45	K	2	B	A		
E			78	M	1	B	A		
F			87	K	1	B	A		
G			3	K	0	C	A		
H			5	K	0	C	A		
I			2	M	0	C	A		
…	…	…	…	…	…	…	…	…	…
K						E	B		
L						E	B		
M						E	B		
N						F	B		
…	…	…	…	…	…	…	…	…	…
P									
Q									
R									

N

GRID CELL A
0 < POP_TOT < N

GRID CELLS
0 < POP_TOT < N

RANK2	GRID-ID	POP_TOT	POP_MAL	POP_FEM	AVE
.. …	A				
.. …	B				
13 224	C				
.. …	..	..	..	..	..
.. …	…	…	…	…	…
.. …	…	…	…	…	…
.. …	…	…	…	…	…
.. …	H				
44 278	I	9	4	5	I
.. …	J				
.. …	…	…	…	…	…
.. …	…	…	…	…	…
.. …	…	…	…	…	…
.. …	N				

lbyrå

Comments to DPRS

- Introduces noise/uncertainty in «all» grid cells
- Noise dependent on population density
- May be reproduced but not disclosed
- Representativity in data, but not for all geographic patterns
- Summarisation, but not exact figures for any area
- May use global hex grid as building block in first step

PORTUGAL Background

- PT has no tradition in applying Statistical Disclosure Control (SDC) methods to grid data

- But we seek to modernize our methods and harmonize them according to the European practice and recommendations

Candidate Methods

Recommended by EU-project
"Harmonized Protection of Census Data in the ESS"

Targeted Record Swapping

- Pre-tabular (applied to microdata)
- Exchange of geographical data between pairs of 'high-risk' 'similar' households

Candidate Methods

Recommended by EU-project
"Harmonized Protection of Census Data in the ESS"

Cell Key Method

- Post-tabular (applied to the table cells)
- Consistently adds unbiased random noise to each table cell

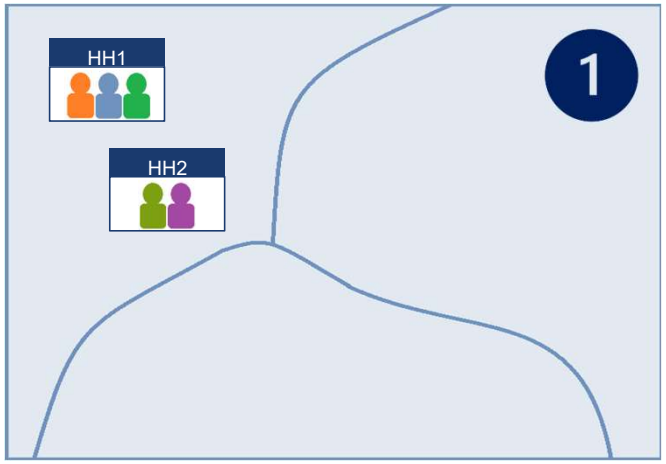
"Countries that do not use a combination of pre and post tabular SDC methods are advised to use the cell key method"

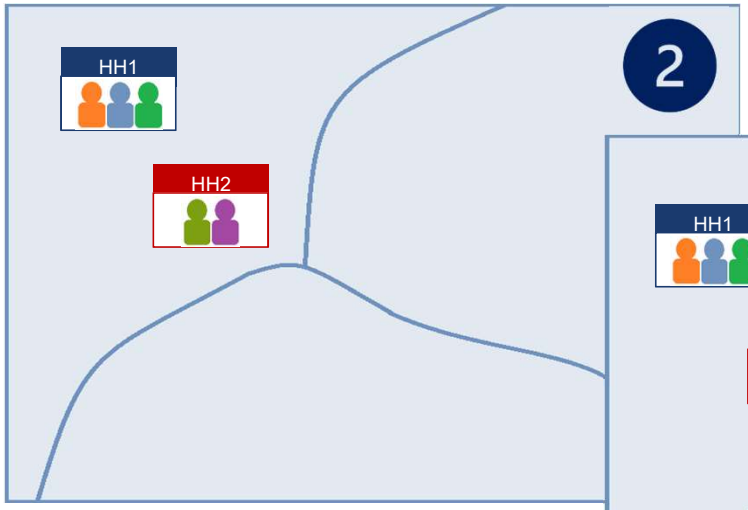
Giessing, S. & Schulte Nordholt, E. (2017) *Recommendations for best practices to protect grid data*. SGA Harmonised protection of census data in the ESS, Work Package 3, Deliverable D3.4

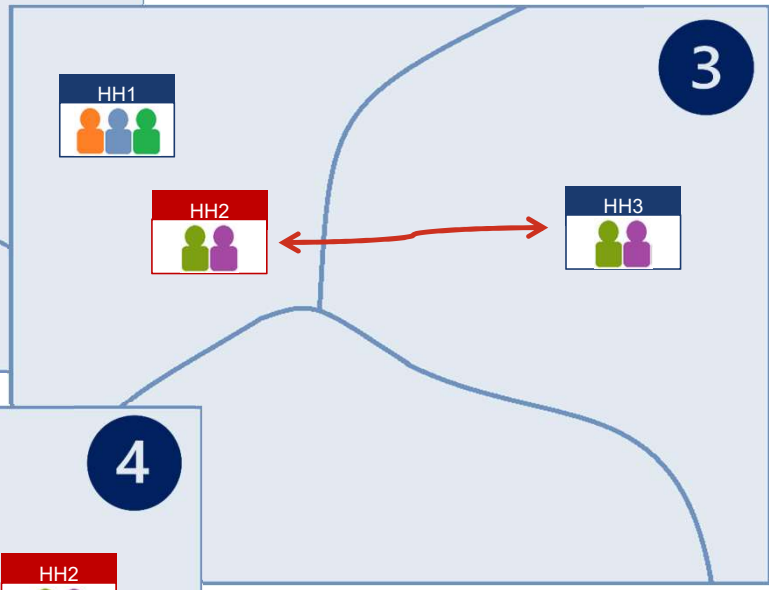
Candidate Methods

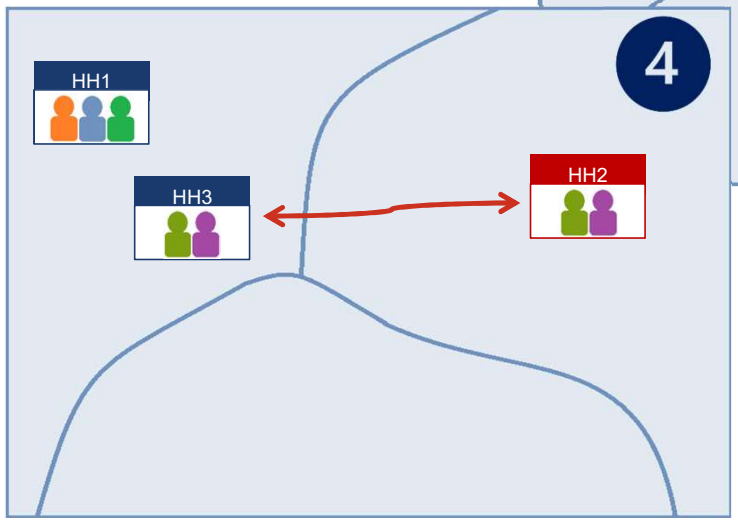
Targeted Record Swapping

(Shlomo et al., 2010)









1. Identify high risk individuals/households
2. Sample high risk households for data swapping
3. Pair the selected households with other households having the same values for the matching variables
4. Swap the geographical information

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

Statistisk sentralbyrå
Statistics Norway

Candidate Methods

Cell key method

(Marley & Leaver, 2011; Enderle et al., 2018)

Microdata

ID	Sex	Age	…	**Record key**
1	1	45	…	**0.13**
2	1	32	…	**0.78**
…	…	…	…	**…**

Frequency table

		Age					
		15-24		25-29		…	
		Count	**Cell key**	Count	**Cell key**	Count	**Cell key**
Sex	1	354	**0.89**	786	**0.24**	…	**…**
	2	632	**0.68**	485	**0.76**	…	**…**
Total		986	0.31	1271	0.53	…	…

Perturbed frequency table

		Age		
		15-24	25-29	…
Sex	1	353	783	…
	2	635	487	…
Total		987	1270	

Perturbation table

		Target frequency			
		0	1	2	…
Original frequency	0	1	0	0	…
	1	0.59	0	0.41	…
	2	0.18	0	0.29	…
	…	…	…	…	…

Candidate Methods

Cell key method

(Marley & Leaver, 2011; Enderle et al., 2018)

- Tested on PT Census 2011 data
- Two groups of EU-hypercubes (Commission Regulation (EU) 2017/712, of 20 April 2017; Commission Implementing Regulation (EU) 2017/543, of 22 March 2017) and some national tables **(no grid data)**
- Risk and utility measures to compare results and support method/parameter choice, **BUT no assessment of disclosure by differencing**

Risks measures

- Let:
  - n_c : number of units that fall into cell c in the original table T
  - n'_c: number of units that fall into cell c in the protected table T'
  - K : total number of cells in table T (or T')

RM 1

Relative change of the number of cells with frequency lower than 3 (change in low frequencies)

$$CLF = \left(\frac{\sum_{c=1}^{K} I(n'_c < 3)}{\sum_{c=1}^{K} I(n_c < 3)} - 1 \right) \times 100\%$$

Instituto Nacional de Estatística
Statistics Portugal

Statistisk sentralbyrå
Statistics Norway

Risks measures

RM 2

Proportion of cells with frequency lower than 3 both in the original and the perturbed table (real low frequencies)

$$RLF = \frac{\sum_{c=1}^{K} I(n_c < 3 \wedge n'_c < 3)}{K} \times 100\%$$

Utility measures

UM 1 (Shlomo & Young, 2005; Shlomo, 2007)

Absolute distance (AD) between the original and the perturbed counts

$$AD_c = |n'_c - n_c|$$

UM 2 (Shlomo & Young, 2005; Shlomo, 2007)

Relative absolute distance (RD) between the original and the perturbed counts

$$RD_c = \frac{|n'_c - n_c|}{n_c}$$

Simple descriptive statistics (max, mean, sd, median) across all cells

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

Statistisk sentralbyrå
Statistics Norway

Utility measures

UM 3 (Buron et al., 2017)

Proportion of false zeros

$$FZ = \frac{\sum_{c=1}^{K} I(n'_c = 0 \wedge n_c \neq 0)}{\sum_{c=1}^{K} I(n'_c = 0)} \times 100\%$$

UM 4

Proportion of unchanged cells

$$UC = \frac{\sum_{c=1}^{K} I(n'_c = n_c)}{K} \times 100\%$$

Challenges

- Disclosure by differencing (e.g. grid cells versus administrative regions) is difficult to measure
- CKM results in loss of table additivity
- CKM can result in false zero frequency cells (but data items on total population shall nevertheless be flagged as 'populated', according to Regulation 1799 of 21 November 2018, Article 6)

Perturbed frequency table

		Age		
		15-24	25-29	…
Sex	1	353	783	…
	2	635	487	…
Total		987	1270	

Instituto Nacional de Estatística
Statistics Portugal

Statistisk sentralbyrå
Statistics Norway

Challenges

Communicating to the users

- Users need to be aware that perturbative SDC methods were used
- Selected disclosure risk and utility indicators might be published together with data
- The loss of table additivity due to confidentiality protection should be clearly stated

Next steps

- Finishing the Census 2021
- Publishing the Data as INSPIRE SERVICE - improving access to data (open data)
- Dissemination of Census 2021 Data

CENSOS 2021

Instituto Nacional de Estatística
Statistics Portugal

Statistisk sentralbyrå
Statistics Norway

Summary

- Lot's of methods, but no single standard for handling confidentiality
- How to handle data in cross border situations?
- Publishing in different grid systems is a challenge
- Could DGGS be an «interchange format» between different grid systems?

Thank You

Ana Santos | Statistics Portugal
Vilni Verner Holst Bloch | Statistics Norway