Disclosure control of Census data by swapping grid-cell identifiers

Steinn Kári Steinsson, Ómar Harðarson, Violeta Calian

Statistics Iceland

14-15 December 2023

Co-funded by the European Union

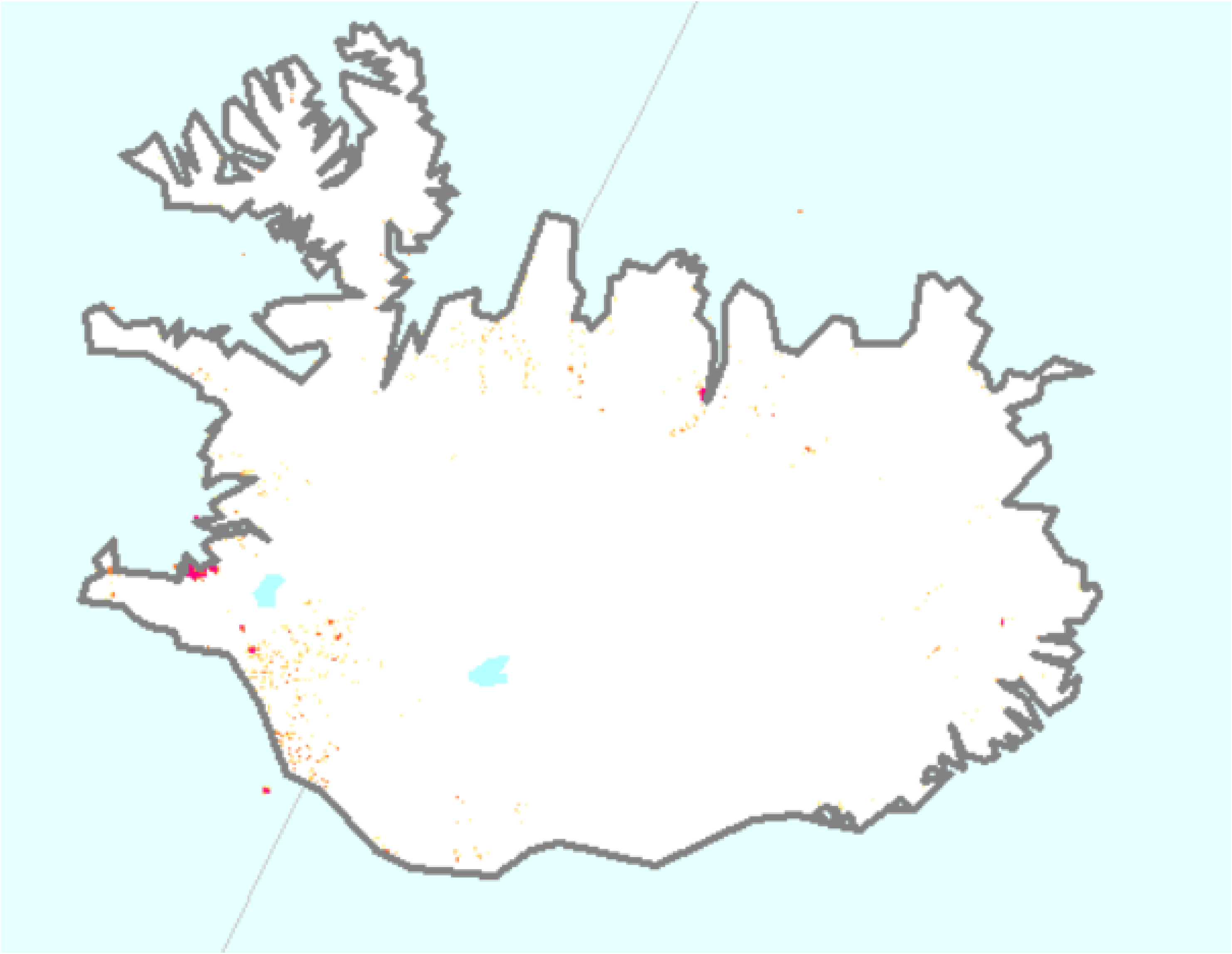
Statistical methods and tools for time series, seasonal adjustment, and statistical disclosure control

Content

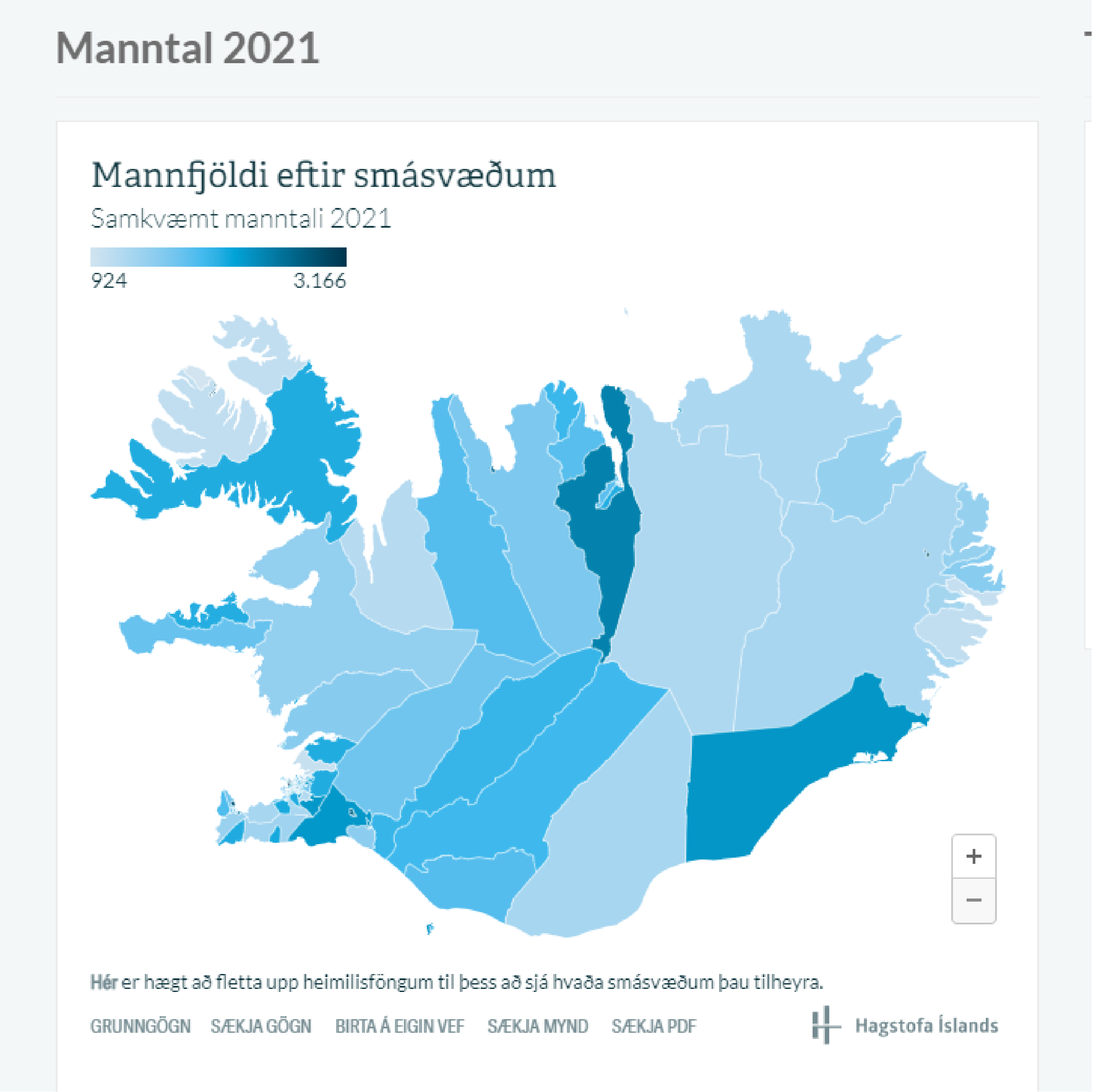
- *Motivation: the case study and the general problem*

- Importance of geospatial dimension
- Research questions
- Method of solution
- Results
- Discussion

https://ec.europa.eu/eurostat/statistics-explained/images/c/c6/GEOSTAT_population_grid_2011.png



Icelandic small output areas



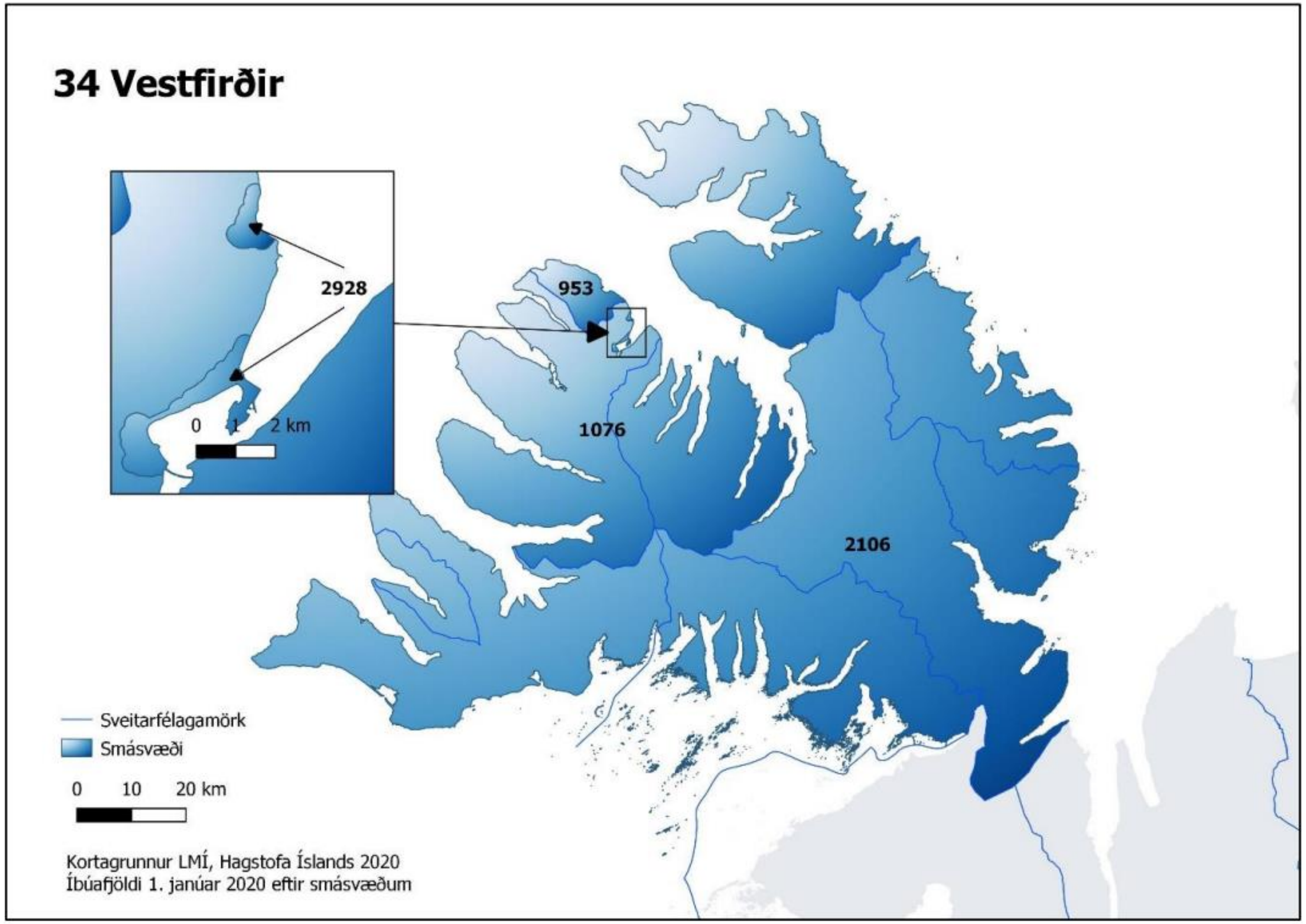
Manntal 2021

Mannfjöldi eftir smásvæðum

Samkvæmt manntali 2021

Hér er hægt að fletta upp heimilisföngum til þess að sjá hvaða smásvæðum þau tilheyra.

Zooming in


34 Vestfirðir
2928
0 1 2 km
953
1076
2106
Sveitarfélagamörk
Smásvæði
0 10 20 km
Kortagrunnur LMÍ, Hagstofa Íslands 2020
Íbúafjöldi 1. janúar 2020 eftir smásvæðum

Data for the 1km2 grid:

https://ec.europa.eu/eurostat/documents/3859598/9670557/KS-GQ-18-010-EN-N.pdf/c3df7fcb-f134-4398-94c8-4be0b7ec0494

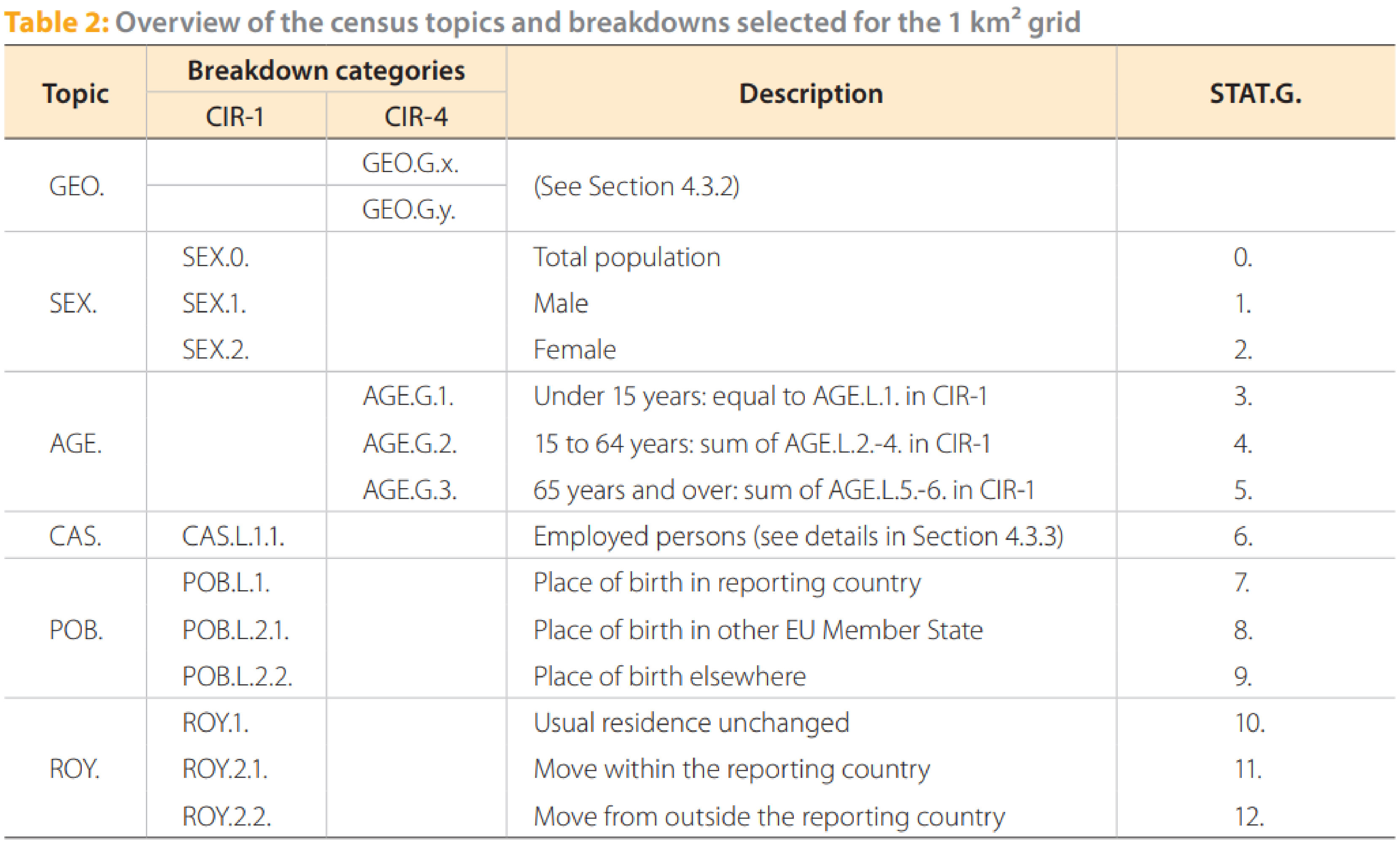
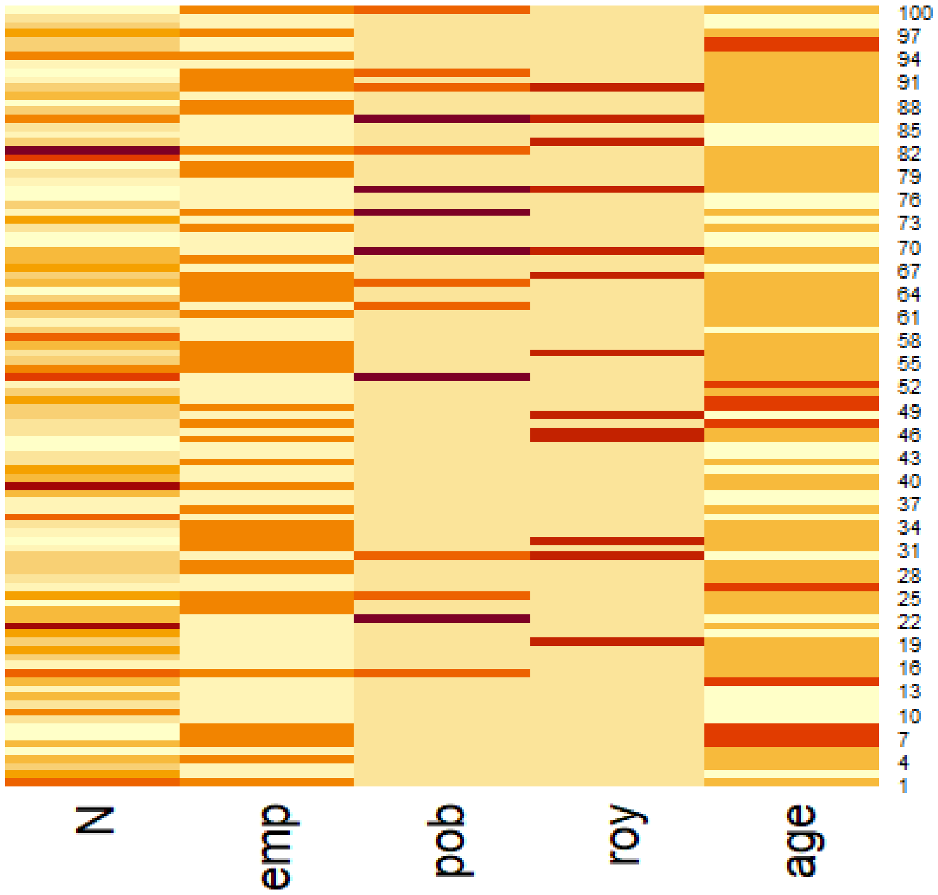


Table 2: Overview of the census topics and breakdowns selected for the 1 km^2 grid

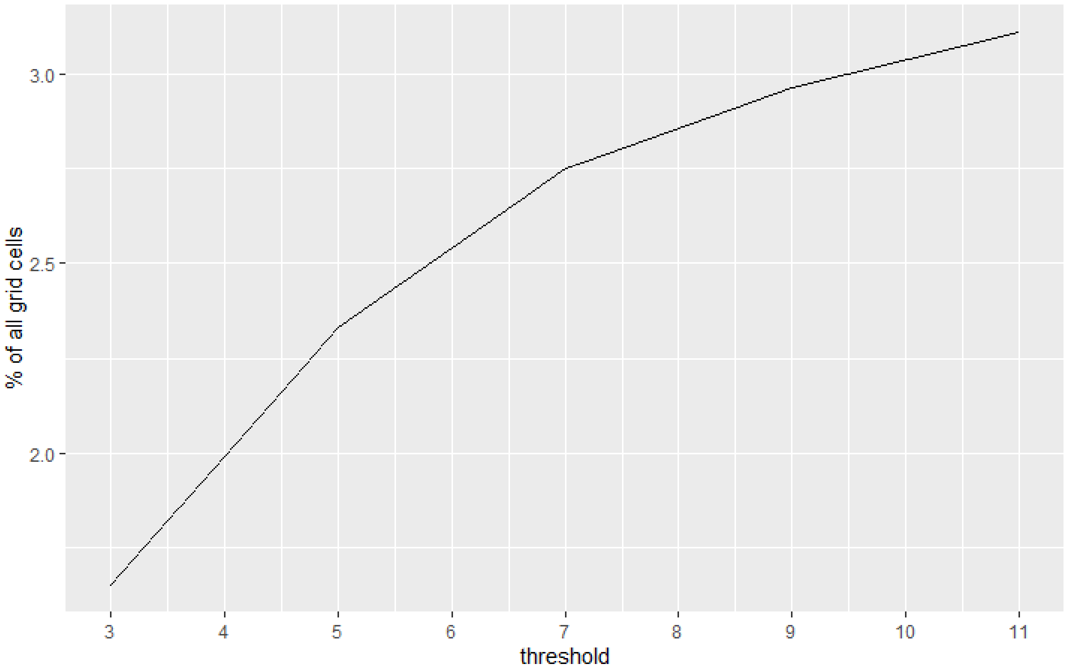
Topic	Breakdown categories		Description	STAT.G.
	CIR-1	CIR-4		
GEO.		GEO.G.x.	(See Section 4.3.2)	
		GEO.G.y.		
SEX.	SEX.0.		Total population	0.
	SEX.1.		Male	1.
	SEX.2.		Female	2.
AGE.		AGE.G.1.	Under 15 years: equal to AGE.L.1. in CIR-1	3.
		AGE.G.2.	15 to 64 years: sum of AGE.L.2.-4. in CIR-1	4.
		AGE.G.3.	65 years and over: sum of AGE.L.5.-6. in CIR-1	5.
CAS.	CAS.L.1.1.		Employed persons (see details in Section 4.3.3)	6.
POB.	POB.L.1.		Place of birth in reporting country	7.
	POB.L.2.1.		Place of birth in other EU Member State	8.
	POB.L.2.2.		Place of birth elsewhere	9.
ROY.	ROY.1.		Usual residence unchanged	10.
	ROY.2.1.		Move within the reporting country	11.
	ROY.2.2.		Move from outside the reporting country	12.

Visualisation: qualitative

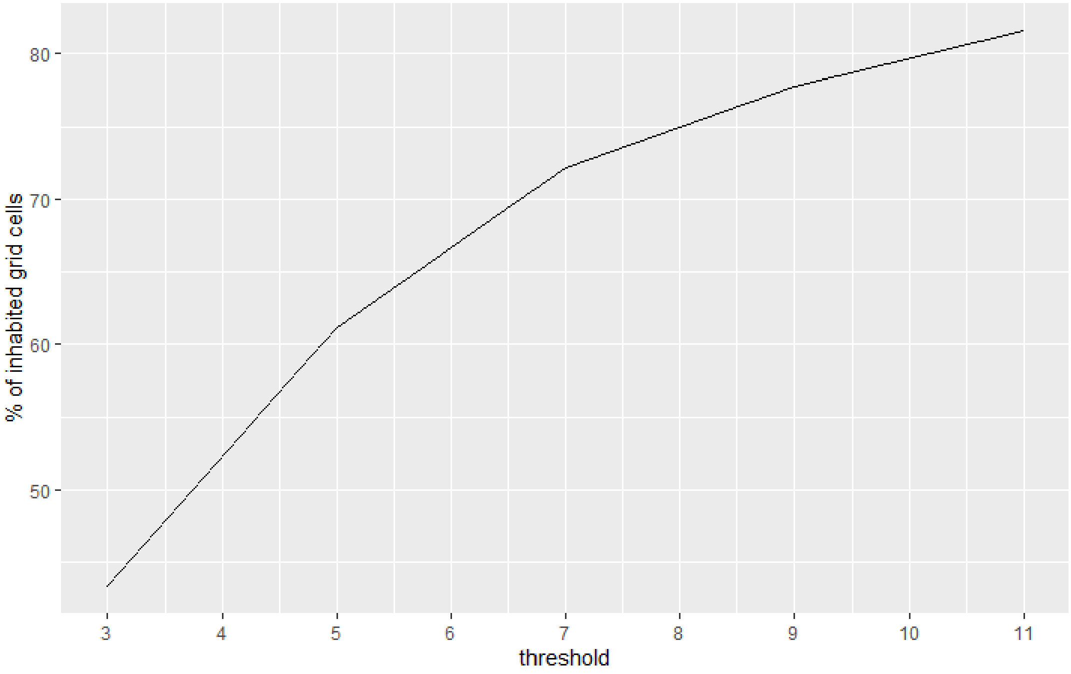


Percentage of permuted cells as a function of threshold

Out of total grid cells



Out of inhabited grid cells



Conditions:

- Conservation of regional (or other spatial scale) distributions

- Credibility:

e.g. conservation of the (approximative) location of special groups which are public knowledge

Research questions

- Q1: does swapping of only cell-ids of “vulnerable” cells provide enough data protection? Should we swap the cell-ids of more cells, e.g. of a percentage of randomly chosen ones?
- Q2: is the cell-id-swapping procedure equivalent/similar to the household-id swapping, differing only by the aggregation level? How do we quantify the similarity-relation?
- Q3: would the newly proposed method provide microdata protection as well?
- Q4: (i) in case Statistics Iceland decides to publish Census results based on more than one grid-system, could we evaluate the disclosure risk increase? (ii) Would two independent cell-id-swapping procedures ensure satisfactory data protection?
- Q5: are there group-identification issues need to be resolved?

Method (more general than needed)

- S1. Microdata (census-grid): D0 -> aggregate* (no cross-tab)-> R0
- S2. Standard record swapping (D0) -> D1 -> aggregate* -> R1
- *S2a. V*ariant: record-swapping (D0) by using grid-cell units instead of household-units -> D2 -> aggregate* -> R2
- S3. Permutation (conditional on similarity profiles and other constraints) of the grid-cell-ids, directly for the tabular version (R0) of the original data -> R3
- S4. Cell-key method -> R4. Added test: whether the hypothesis of no association between the (very) small areas and perturbation count variables can/not be rejected
- S5. Risk and Utility measures for R0-R4 obtained for varying thresholds involved in defining these steps. Comparing but not optimizing!
- S6. One more system of small areas! In order to investigate whether publishing both grids can lead to differencing disclosure issues. It consists of the areas belonging to both grids and the areas belonging only to one or the other. If the cell-ids of the Icelandic-grid and EU-grid are permuted *independently*, the resulted system has a very low differencing disclosure risk.

Preliminary results

- Q1: (target) swapping of *cell-ids* provides data protection, for ***this*** problem
- Q2: the cell-id-swapping procedure ~related to~ the household-id swapping but application may be done at D0 or *R0* levels.
- Q3: microdata protection by cell-id swapping? Extra-analysis on.
- Q4: if Statistics Iceland decides to publish Census results based on more than one grid-system: two independent cell-id-swapping procedures ensure satisfactory data protection
- Q5: group-identification issues to be resolved (~)

Discussion

- Difficulties (definitions/measures, conventions, conditions)
- General goal: to provide data protection
  - in a *systematic* manner and
  - as *automatic* as possible.
- New ideas to investigate:
  - using Bayesian estimates instead of raw counts (*experimented in a particular context)
  - using deep-learning/cryptography inspired methods.
- Sharing the code:

  https://github.com/violetacln/testingSDCtools

References

[1] Hundepool, A., DomingoFerrer, J., Franconi, L., Giessing, S, Nordholt, E. S., Spicer, K., de Wolf, P.P., Shewhart, A., Wilk, S., Statistical Disclosure Control, Wiley, 2012.

[2] Shlomo, N., Statistical Disclosure Control Methods for Census Frequency Tables, International Statistical Review / Revue Internationale de Statistique, 75(2), 199-217, 2007.

[3] Templ M, Kowarik A, Meindl B (2015). “Statistical Disclosure Control for Micro-Data Using the R Package sdcMicro.” *Journal of Statistical Software*, **67**(4), 1–36. doi:10.18637/jss.v067.i04.

[4] Methods of statistical disclosure control for aggregate data. With a case study on the new Icelandic geospatial system of statistical output areas, Violeta Calian. Statistical series: Working papers, 105(6), 2 September 2020

[5] Benschop, T., Machingauta, C., Welch, M. (2022) Statistical Disclosure Control: A Practice Guide, https://buildmedia.readthedocs.org/media/pdf/sdcpractice/latest/sdcpractice.pdf

[6] Shlomo, N., Antal, L. and Elliot, M. (2013) Measuring Disclosure Risk and Data Utility for Flexible Table Generators, Joint UNECE/Eurostat work session on statistical data confidentiality, https://unece.org/sites/default/files/2022-11/ECECESSTAT20226.pdf

Thank you!

violeta.calian@hagstofa.is