Population grids protected by the Cell Key Method

Martin Möhler

Federal Statistical Office (Destatis)

Question

Population data for Census 2022 geographic grids (1ha) in Germany and several other nations will be protected by the Cell Key Method, a post-tabular perturbation technique [1]. Is the fitness for purpose of the perturbed raster preserved? To address the question I run three archetypal spatial analysis tasks on a confidential (unprotected) raster and its perturbed counterparts (two variants) and compare the results.

Cell Key Method

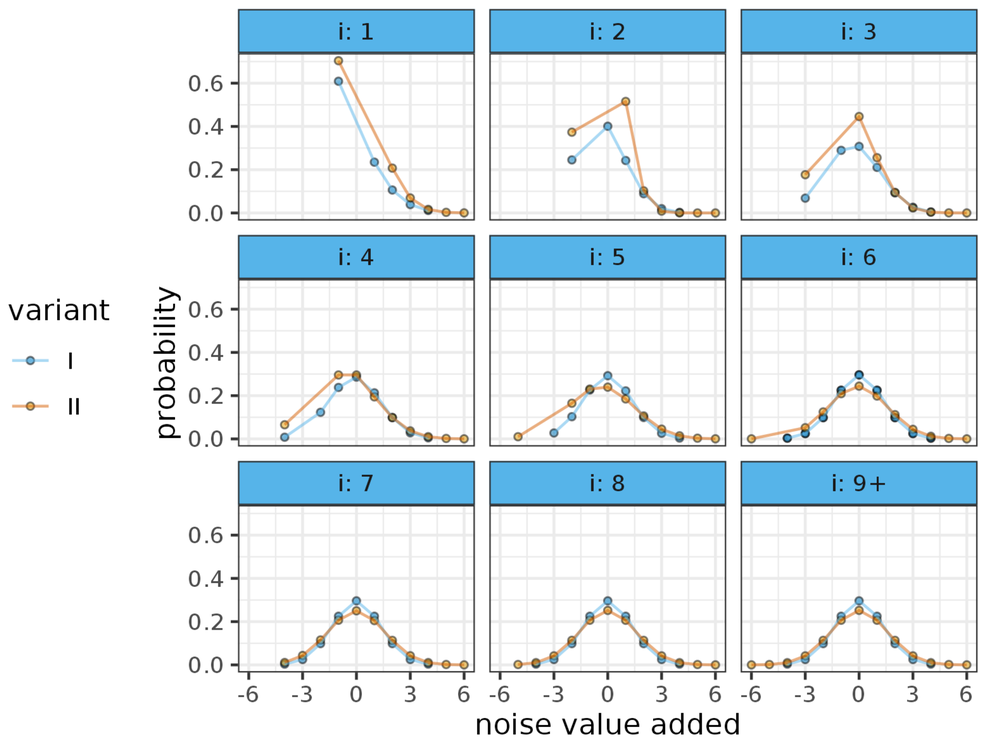
Perturbed grid value: $j = i + \Delta i$ with noise distribution $P_i(\Delta i \mid i)$ [2], [3]:

$$\max_{\Delta i} \; -\sum P_i(\Delta i \mid i) \log\left[P_i(\Delta i \mid i)\right] \;,\; i \in \mathbb{N}$$

$$\text{s.t. } |\Delta i| \leq D,\; E(\Delta i \mid i) = 0,\; \mathrm{Var}(\Delta i \mid i) = \sigma^2,\; j \geq 0,\; j \notin \{1, \ldots, j_s\}$$

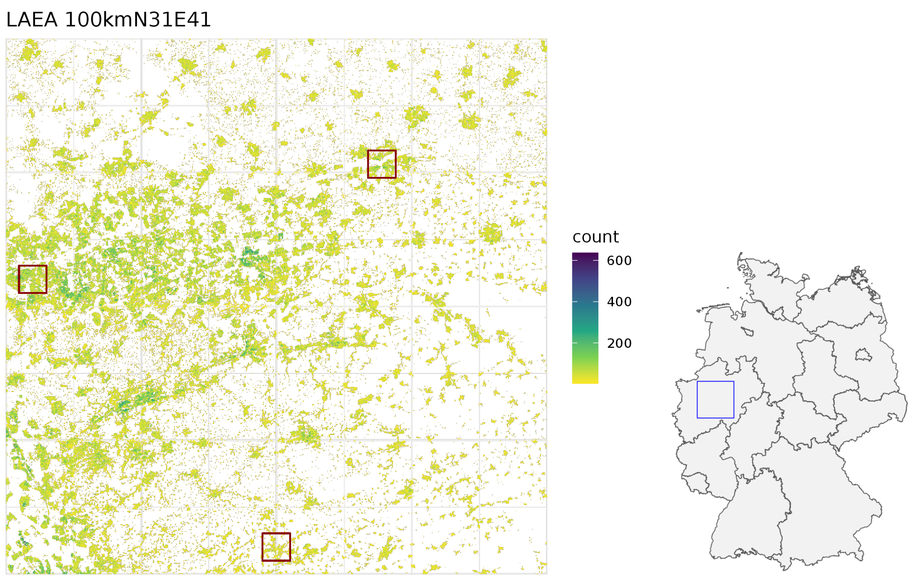
D max. abs. perturbation
σ perturbation variance
j_s small counts threshold

Variant I: $D = 4, \sigma = 1.8, j_s = 1$
Variant II: $D = 6, \sigma = 2.5, j_s = 2$



Data

Census 2011 population counts over 100km $\times$ 100km area in North Rhine - Westphalia, Germany; full map: 1 mio. cells à 100m $\times$ 100m



For *local* analyses I consider 3 focus areas (25km^2, 50 $\times$ 50 cells):

- *area 1* (west): central urban, high density
- *area 2* (north-east): mixed urban / suburban, medium density
- *area 3* (south): rural, low density

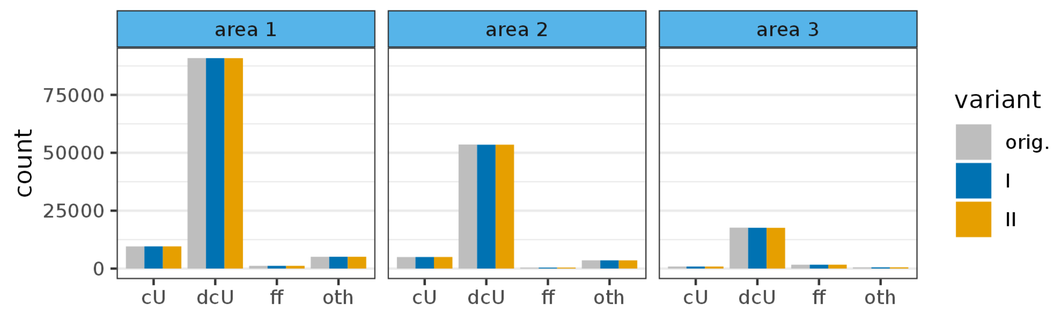
Data sets for comparison: original grid, grid with Cell Key Method variant I (weaker), grid with Cell Key Method variant II (moderate)

Data sources:

Grid-level population counts: Statistisches Bundesamt, Destatis
(https://www.zensus2022.de/DE/Was-ist-der-Zensus/gitterzellenbasierte_Ergebnisse_Zensus2011.html)
CORINE Land Cover 5ha 2018: Bundesamt für Kartographie und Geodäsie (BKG)
(https://gdz.bkg.bund.de/index.php/default/open-data/wfs-corine-land-cover-5-ha-stand-2018-wfs-clc5-2018.html)

Analysis 1: Raster overlay

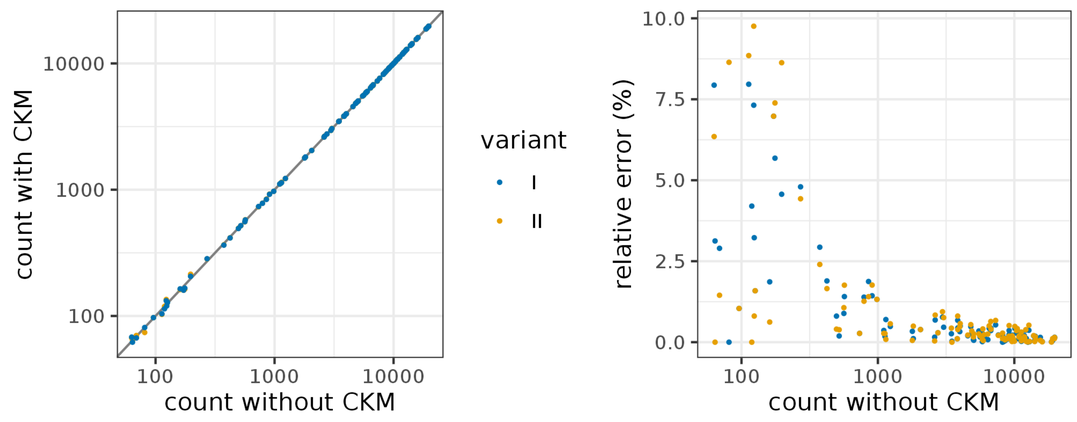
- Task: Classify and aggregate by land cover type
- Context: Environmental impact analysis, quality of life indices
- Comparison: Population coincidence by land cover type



cU: continuous urban, dcU: discontinuous urban, ff: fields and forests, oth: other

Analysis 2: Buffer operations

- Task: Sum population counts over buffer zones
- Context: Proximity to utilities, exposure to environmental effects
- Comparison: 100 randomly placed circular zones (à 1km radius)



Analysis 3: Hot spots

- Task: Determine spots where pop. > 80% quantile
- Context: Urban planning, crime & disease prevention
- Comparison: Hot spot overlap [4] (*left table*) and population coincidence in hot spots (*right table*)

Jaccard similarity (100 = best)		
map	variant I	variant II
full	97.5	97.0
area 1	97.6	97.3
area 2	96.4	96.0
area 3	93.4	92.9

Pop. in hotspots (% difference)		
map	variant I	variant II
full	0.07	0.08
area 1	1.41	1.63
area 2	-0.81	-1.06
area 3	-2.93	-2.53

Conclusion

Population grids protected by the Cell Key Method overall preserve analytical validity for standard spatial analysis tasks. Stronger perturbation implies slightly more loss of accuracy. Sparsely populated areas are impacted more. Increasing the number of grid cells considered leads to a more robust result.

References

[1] M. de Vries, M. Golmajer, R. Tent, S. Giessing, and P.-P. de Wolf, "An overview of used methods to protect the European Census 2021 tables", in *UNECE Expert Meeting on Statistical Data Confidentiality, September 26-18, Wiesbaden, Germany*, 2023.

[2] J. K. Marley and V. L. Leaver, "A method for confidentialising user-defined tables: Statistical properties and a risk-utility analysis", in *Proceedings of the 58th World Statistical Congress, Dublin*, 2011.

[3] S. Giessing, "Computational issues in the design of transition probabilities and disclosure risk estimation for additive noise", in *UNESCO Chair in Data Privacy International Conference, PSD '16, Dubrovnik, Croatia, Proceedings*, 2016.

[4] P.-P. de Wolf and E. de Jonge, "Location related risk and utility", in *UNECE Work Session on Statistical Data Confidentiality, September 20-22, Skopje, North Macedonia*, 2017.

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
Finanziert von der Europäischen Union
NextGenerationEU