# Spatial SDC experiments and evaluations with multiple countries comparison

Johannes Gussenbauer (STAT), Julien Jamme (INSEE), Edwin de Jonge (CBS),
Peter-Paul de Wolf (CBS), Martin Möhler (Destatis)

UNECE Expert meeting on Statistical Data Confidentiality 2023
26.09.2023-28.09.2023 Weisbaden, Germany

This work is part of the Centre of Excellence on SDC (CoE on SDC) and is co-funded by the European Commission by means of grant agreement number 899218, 2019-BG-Methodology.

Co-funded by
the European Union

# Overview

- Experiment Setup

- Risk and Utility Measures

- Results and Discussion

# Experiment Setup

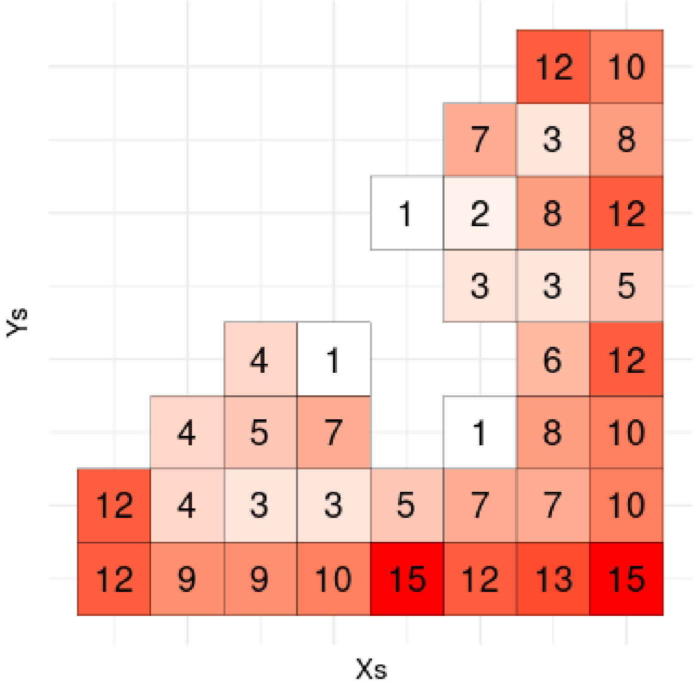
- ‘Census-like’ datasets from 4 countries
  - Austria, France, Germany, Netherlands

- Tabular data on person level with coordinates of residence and grid cells
  - INSPIRE2014 standard ETRS89-LAEA

- Aim: test and compare several methods for protecting grid data

# ‘Census-like’ dataset and map

| Person ID | Grid cell ID | Ys | Xs | … |
|---|---|---|---|---|
| 00000001 | 500mN28215E46275 | 2821500 | 4627500 | … |
| 00000002 | 500mN28085E47890 | 2808500 | 4789000 | … |
| 00000003 | 500mN28025E47925 | 2802500 | 4792500 | … |
| 00000004 | 500mN28120E47985 | 2812000 | 4798500 | … |
| … | … | … | … | … |



# Experiment Setup

1. Build table on count data (~number of people) by grid cells (L000500)

2. Calculate risk measures

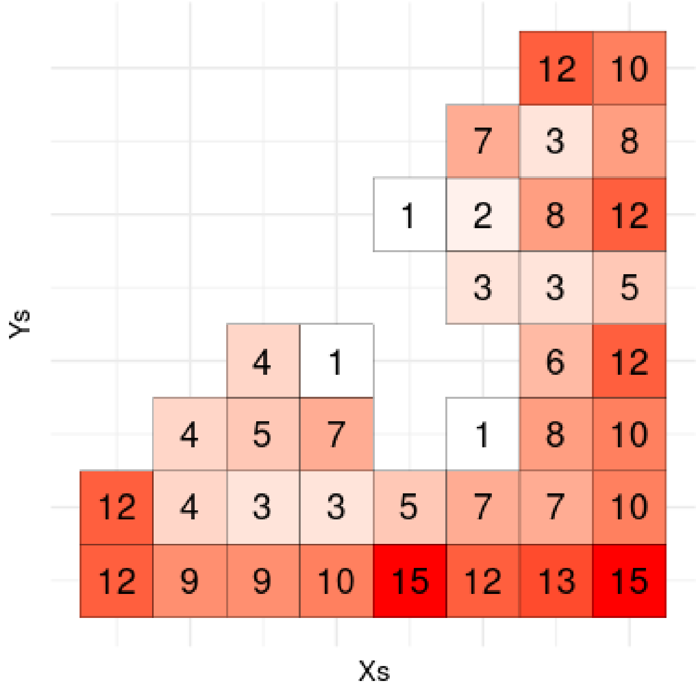
3. Apply SDC methods using the R-Package `sdcSpatial`

4. Re-evaluate risk measures and calculate information loss

# Protection Methods

- **Cell removal**: suppresses the sensitive cell

Original map

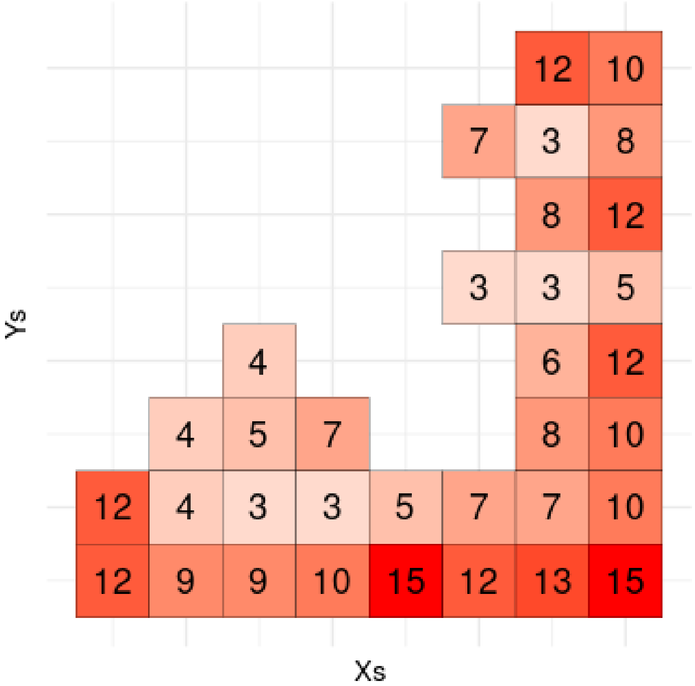


| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | 12 | 10 |
| | | | | | 7 | 3 | 8 |
| | | | | 1 | 2 | 8 | 12 |
| | | | | | 3 | 3 | 5 |
| | | 4 | 1 | | | 6 | 12 |
| | 4 | 5 | 7 | | 1 | 8 | 10 |
| 12 | 4 | 3 | 3 | 5 | 7 | 7 | 10 |
| 12 | 9 | 9 | 10 | 15 | 12 | 13 | 15 |

# Protection Methods

- **Cell removal**: suppresses the sensitive cell

Protected map



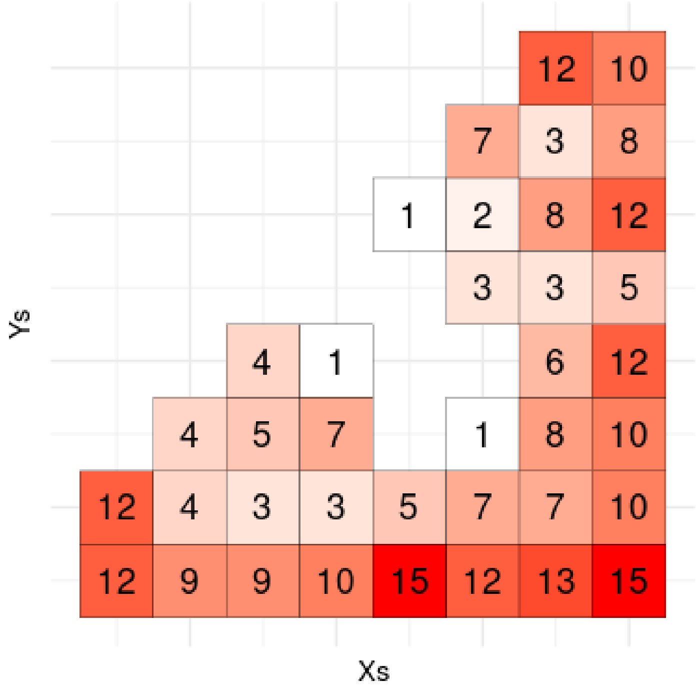
| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | 12 | 10 |
| | | | | | 7 | 3 | 8 |
| | | | | | | 8 | 12 |
| | | | | | 3 | 3 | 5 |
| | | 4 | | | | 6 | 12 |
| | 4 | 5 | 7 | | | 8 | 10 |
| 12 | 4 | 3 | 3 | 5 | 7 | 7 | 10 |
| 12 | 9 | 9 | 10 | 15 | 12 | 13 | 15 |

# Protection Methods

- **Quad tree**: aggregate sensitive cells with its three neighbours

- Can *zoom-out* multiple times

Original map

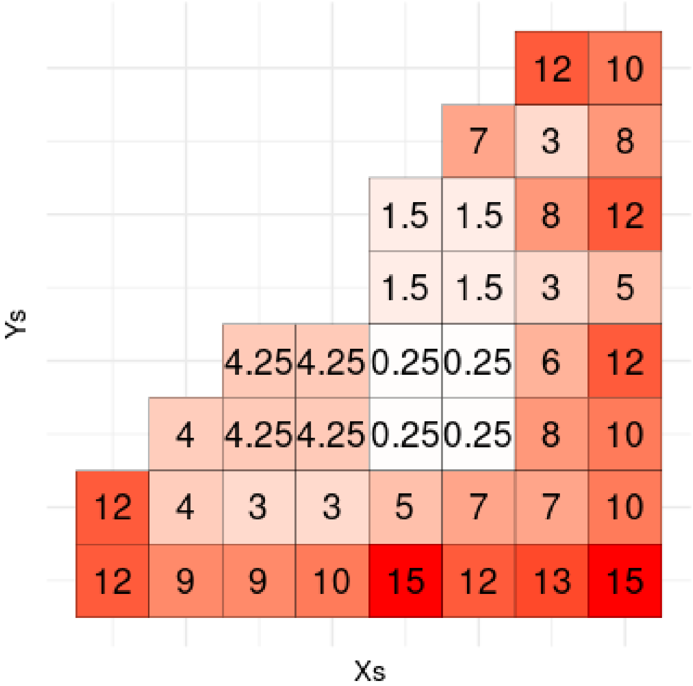


# Protection Methods

- **Quad tree**: aggregate sensitive cells with its three neighbours

- Can *zoom-out* multiple times

Protected map



| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | 12 | 10 |
| | | | | | 7 | 3 | 8 |
| | | | | 1.5 | 1.5 | 8 | 12 |
| | | | | 1.5 | 1.5 | 3 | 5 |
| | | 4.25 | 4.25 | 0.25 | 0.25 | 6 | 12 |
| | 4 | 4.25 | 4.25 | 0.25 | 0.25 | 8 | 10 |
| 12 | 4 | 3 | 3 | 5 | 7 | 7 | 10 |
| 12 | 9 | 9 | 10 | 15 | 12 | 13 | 15 |

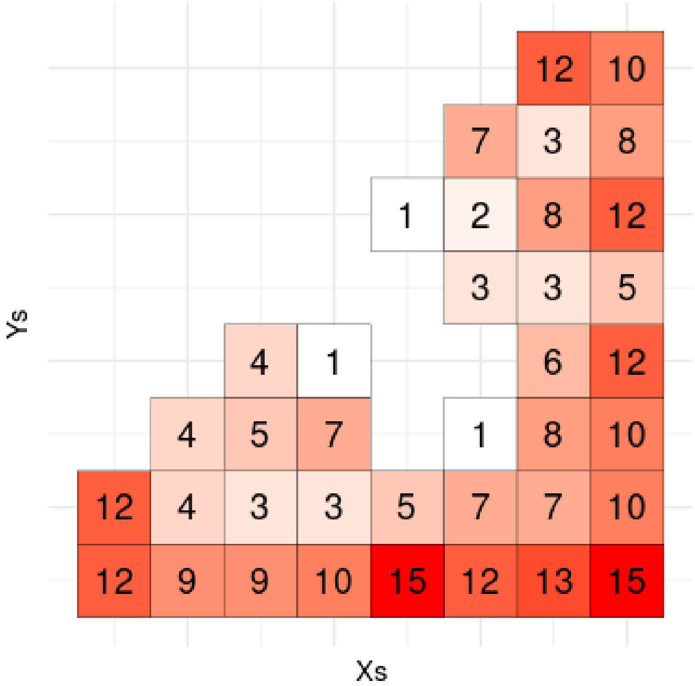
# Protection Methods

- **Kernel density smoothing**: mass of population is spread out over a neighbouring region

$$\hat{f}_h(x,y) = \frac{1}{h^2}\sum_{i=1}^{N} K\left(\frac{x-x_i}{h}, \frac{y-y_i}{h}\right)$$

$K(x,y)$ bivariate Gaussian kernel

Original map



| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | 12 | 10 |
| | | | | | 7 | 3 | 8 |
| | | | | 1 | 2 | 8 | 12 |
| | | | | | 3 | 3 | 5 |
| | | 4 | 1 | | | 6 | 12 |
| | 4 | 5 | 7 | | 1 | 8 | 10 |
| 12 | 4 | 3 | 3 | 5 | 7 | 7 | 10 |
| 12 | 9 | 9 | 10 | 15 | 12 | 13 | 15 |

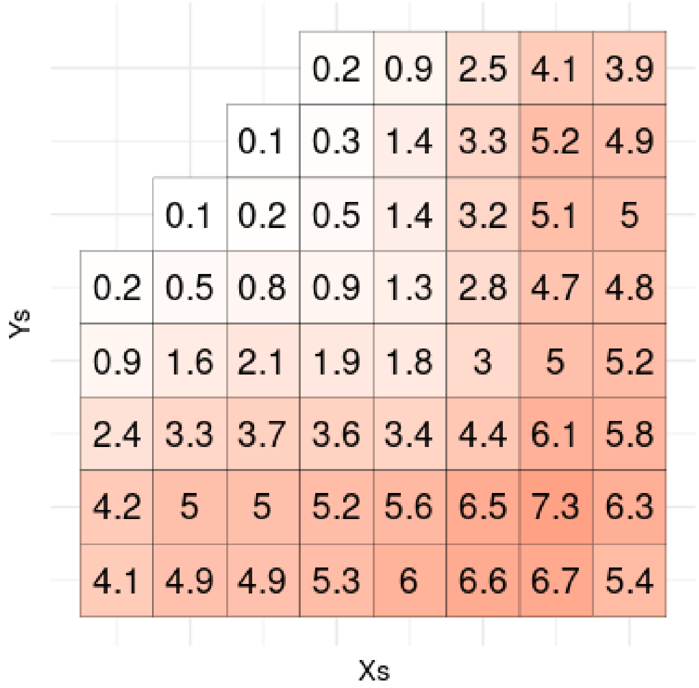
# Protection Methods

- **Kernel density smoothing**: mass of population is spread out over a neighbouring region

$$\hat{f}_h(x,y) = \frac{1}{h^2}\sum_{i=1}^{N} K\left(\frac{x-x_i}{h},\frac{y-y_i}{h}\right)$$

$K(x,y)$ bivariate Gaussian kernel

Protected map



| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | 0.2 | 0.9 | 2.5 | 4.1 | 3.9 |
| | | 0.1 | 0.3 | 1.4 | 3.3 | 5.2 | 4.9 |
| | 0.1 | 0.2 | 0.5 | 1.4 | 3.2 | 5.1 | 5 |
| 0.2 | 0.5 | 0.8 | 0.9 | 1.3 | 2.8 | 4.7 | 4.8 |
| 0.9 | 1.6 | 2.1 | 1.9 | 1.8 | 3 | 5 | 5.2 |
| 2.4 | 3.3 | 3.7 | 3.6 | 3.4 | 4.4 | 6.1 | 5.8 |
| 4.2 | 5 | 5 | 5.2 | 5.6 | 6.5 | 7.3 | 6.3 |
| 4.1 | 4.9 | 4.9 | 5.3 | 6 | 6.6 | 6.7 | 5.4 |

# Risk and Utility Measures

- Grid cell $\mathcal{C}_j$ is at risk if it contains fewer than $k$ people
- Risk measure ~ share of grid cells/population which are *at risk*

$$R^{(\mathcal{C})}(k) := \frac{1}{M} \sum_{j=1}^{M} R_j\,(k) \qquad\qquad R^{(N)}(k) := \frac{1}{N} \sum_{j=1}^{M} R_j\,(k) \cdot r_j$$

with

$$R_j(k) = \mathbb{I}\left[r_j < k\right] \quad \forall j = 1, \dots, M$$

$$r_j = \sum_{i=1}^{N} \mathbb{I}\left[(x_i, y_i) \in \mathcal{C}_j\right] \qquad \forall j = 1, \dots, M$$

$(x_i, y_i)$ coordinates of person $i$

# Risk and Utility Measures

- (normalised) Hellinger's distance between raster $\mathbf{R}$ and $\mathbf{R}'$

$$HD(\mathbf{R},\mathbf{R}') = \frac{1}{\sqrt{2}} \sqrt{\sum_{j=1}^{M} \left( \sqrt{\frac{r'_j}{\sum_{j=1}^{M} r'_j}} - \sqrt{\frac{r_j}{\sum_{j=1}^{M} r_j}} \right)^2}$$

- Easy calculation
- Applicable to tabular data

- Does not account for spatial distribution

# Risk and Utility Measures

- Kantorovic-Wasserstein Distance (KWD) or *Earth Mover Distance*
- Minimal cost to transport a mass from one distribution to another

Shift around distribution mass of $\Delta r_{jk}$ between the $j$th and $k$th grid cell, until $\mathbf{R}'$ is transformed into $\mathbf{R}$

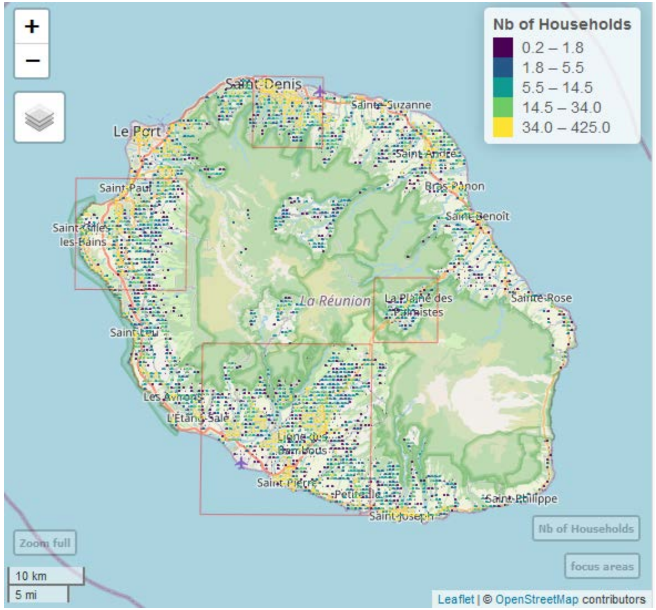
- Considers spatial distribution
- Intuitive interpretation

- Difficult to compute ~ R-Package `SpatialKWD`
- Needs methodological choices
  - How to deal with different mass in $\mathbf{R}'$ and $\mathbf{R}$
  - Focus Area
  - Convex hull true/false

# Results

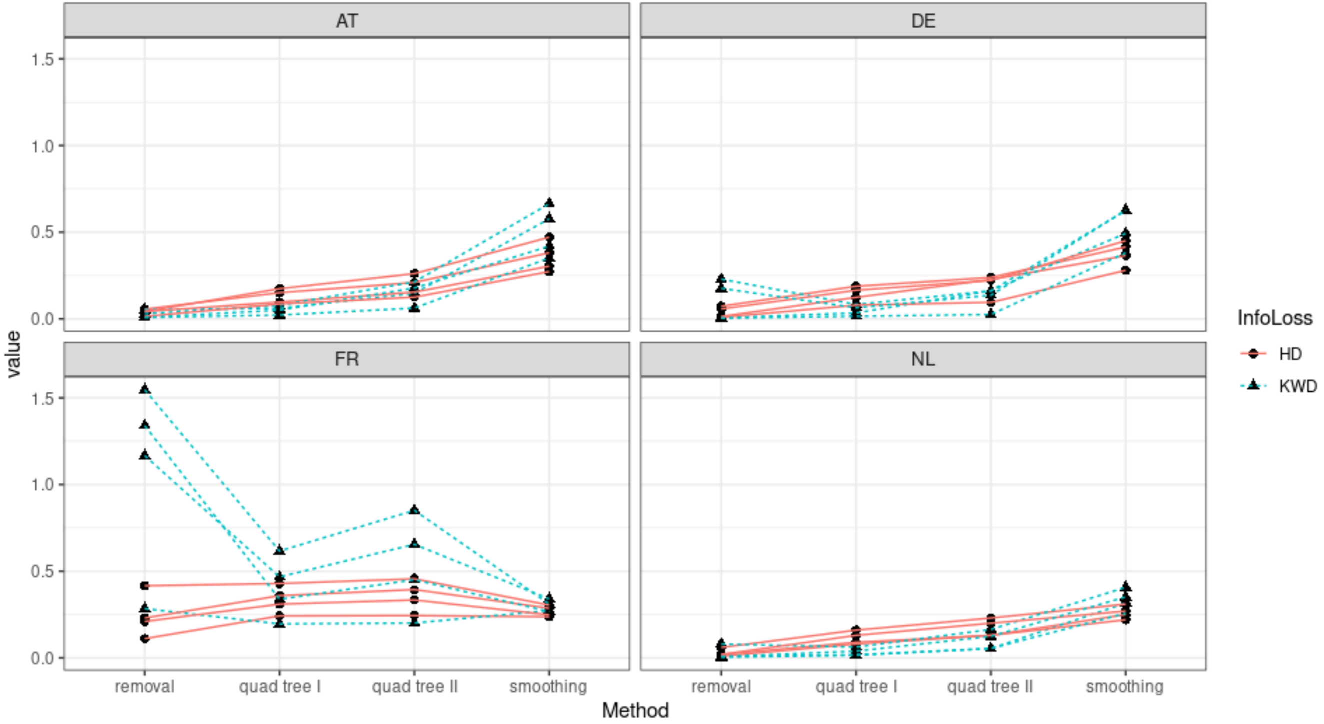
- Depending on the country slightly different setup
  - CBS, DESTATIS, STAT: 500m × 500m, $k = 5$
  - INSEE: 200m × 200m, $k = 11$

- Each country selected **4** specific focus areas
  - Protection applied on whole data set beforehand → focus in on area of interest
    - Can deal with mass missmatch
  - Focus areas contain different population distributions
  - Homogeneously populated, *hot spots*, country borders and uninhabitable terrain.



Island of La Réunion, use case INSEE; Red squares are the focus areas

# Results



AT
DE
FR
NL
value
1.5
1.0
0.5
0.0
removal
quad tree I
quad tree II
smoothing
Method
InfoLoss
HD
KWD

# Conclusions and Discussion

- Cons and Pros of protection methods

| Method | Pros | Cons |
|---|---|---|
| Cell removal | No artificially inhabitable cells | low density regions might be deleted |
| | hot spots kept intact | reidentification risk through differencing |
| Quadtree | easy to apply | overly blocky structure |
| | utility loss rather small | can enlarge hot spots |
| | | can populate uninhabitable cells |
| Smoothing | Hot spots are usually kept intact | Applied to whole data |
| | | can populate uninhabitable cells |

# Conclusions and Discussion

- HD and KWD usually rank methods similar
  - Protection was only applied very locally
  - Some methodological choices needed before applying KWD
    - Impact of different specifications needs more investigation
  - Looking at focus areas instead of whole country more insightful

- Possible additions/improvements to sdcSpatial:
  - Respect borders or natural barriers during protection

- Further analysis needed for
  - Differencing attacks
  - Compare more utility measures (Moran's I, Spatial K-function, Hotspot preservation, preservation of population by type of land cover, …)
  - Compare with more *classical* methods like record swapping or cell key

Code for running experiment on dummy data on github: https://github.com/sdcTools/sdcSpatialExperiment