Map protection methods, introducing MRA Adaptive Maps

Edwin de Jonge @edwindjonge

Statistics Netherlands Research & Development
@edwindjonge

October 7, 2024

Conference on Anonymization of Integrated and Georeferenced Data

Enhancing privacy protected maps

Who am I?

- Methodologist / Data Scientist @cbs.nl / Statistics NL
- Statistics Netherlands (CBS) is producer main official statistics in the Netherlands:
  - Stats on Demographics, economy (GDP), education, environment, agriculture, Finance etc.
  - Part of the European Statistical System, ESS
  - Like DEStatis, only national office (smaller country. . .)

Some Aspects Geographical confidentiality

- Legal, what is allowed?
- Cultural, what is considered sensitive?
- Computational (methods)
- What kind of Information?
- Combining sources with spatial information
- Analysing geo-referenced data, who has access?
- Publishing geo-referenced data

Side note

Country Context may be different.

- The privacy of citizens must be protected, EU/GDPR. (good thing!)
- Sensitive information of citizens should be not be disclosed.
- What is sensitive is not an absolute given:
  - Scandanavian countries had a public register of income.
  - The Netherlands has a public register of buildings, locations and their addresses. who is residing where is not public.
- Tension between privacy and public safety. In times of crisis... (i.e. COVID19 pandemic)

This presentation

- Legal
- Cultural
- Computational
- kind of information
- Combining information with spatial information
- Analysing geo-referenced data, who has access?
- **Publishing geo-referenced data on maps**
- Tension between privacy and public safety. In times of crisis... (i.e. COVID19 pandemic)

This presentation

- Being a statistical office, we do not disclose information of individuals.
- Main task is to compile information on population level.
- **Will focus on methods of publication of (density) maps**
- Focus on publishing regional data, i.e. on the population living or operating in the spatial dimension.
- e.g. Unemployment, or Spatial distribution of income etc.
- So we are talking about "aggregated" data: and the protection takes place at an aggregate level, not on a microlevel. Which is a different topic.
- Assumption: the raw / original data has most information / utility. (come back later)

This presentation

- Uses material from Guidelines for Statistical Disclosure Control Methods Applied on Georeference Data, (Möhler et al. 2024) (to be published)
- contains examples with R package: `sdcSpatial`
- Describes methods and their aspects
- Introduces a new adaptive wavelet method.

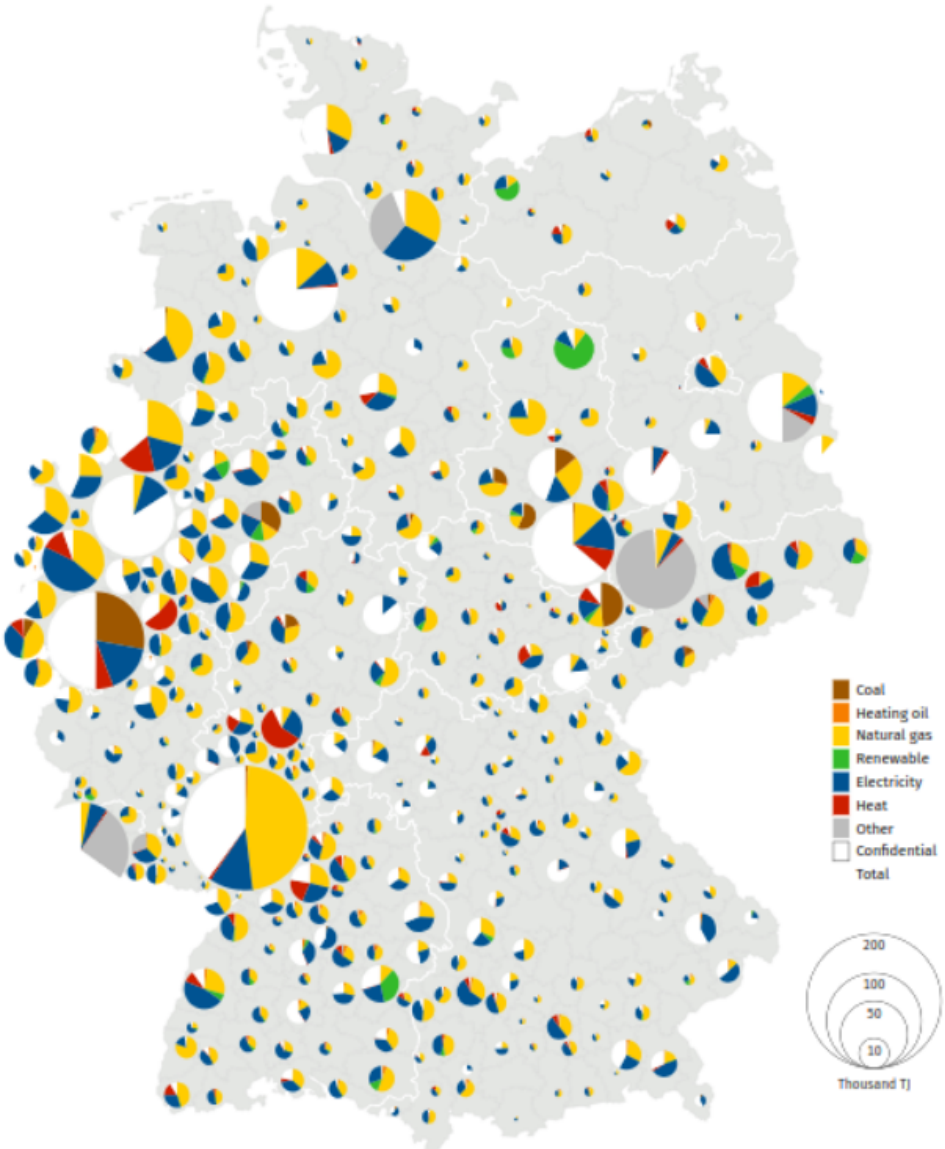
Publishing maps

- Demand of policymakers and society to have ever more detailed spatial detail.
- Often on administrative regions
- More detail, more risk for disclosing individual information...

Maps for a long time:

- Attractive and useful means for displaying spatial data (maps earlier then charts...)
- Maps may disclose information on individuals
- Different type of maps: Proportional Symbol Maps, Choropleths, Density maps

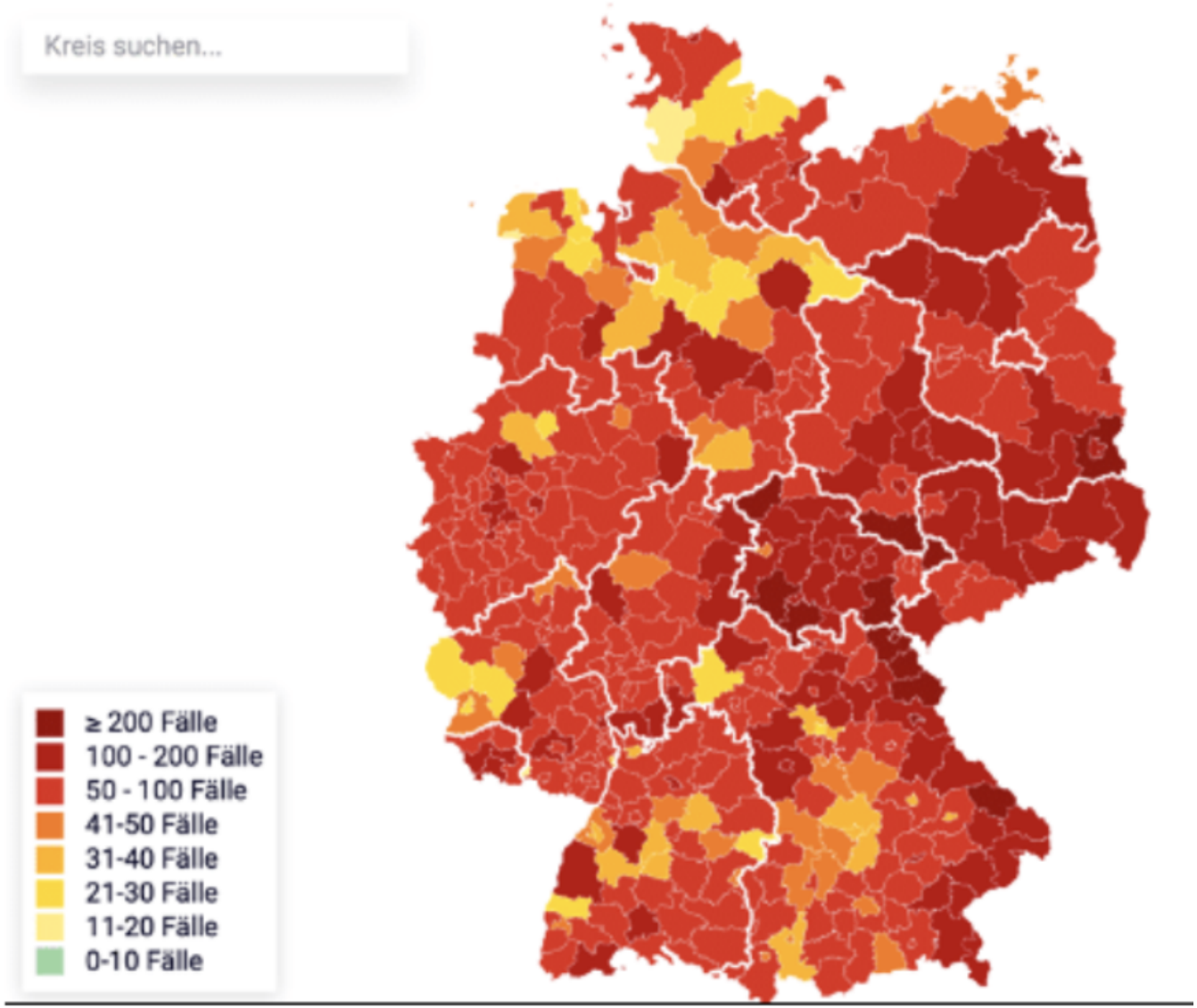
Map: proportional symbol



Map: proportional symbol:

- Used for a long time (1858, Minard)
- Publish **absolute numbers**
- Can be multi variate
- Mostly on administrative regions
- Privacy protection methods: same as tabular methods, “it’s just a table”.
- *Spatial dimension (mostly) not used*

Maps: choropleth:



Maps: choropleth:

- Choropleth: **relative numbers** (percentage/density)
- Mostly univariate
- Mostly on administrative regions
- Privacy protection methods: same as tabular methods, “it’s just a table”.
- *Spatial dimension (mostly) not used*

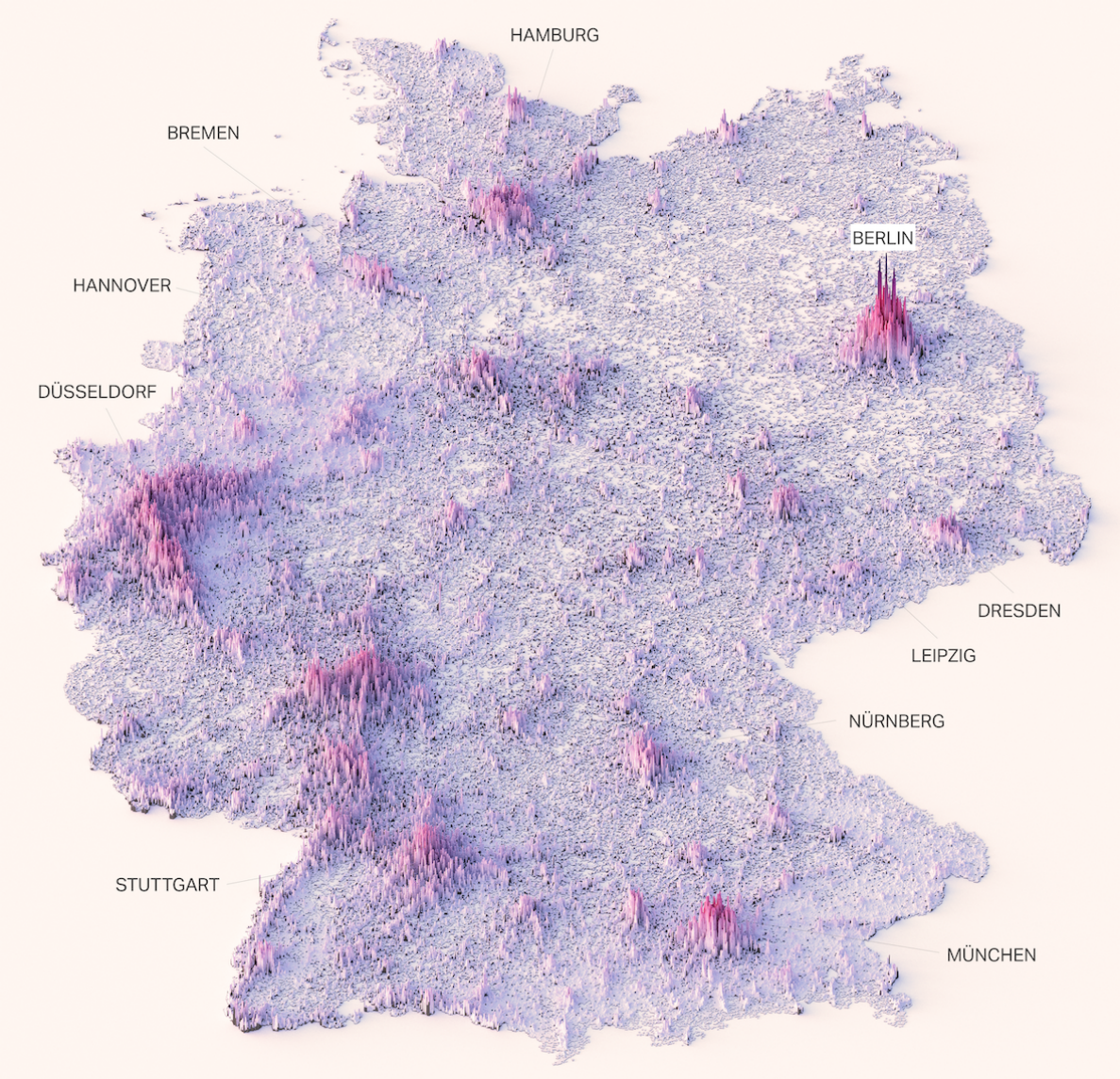
Adminstrative regions

- Very important for policy makers!

but have spatial distortion

- e.g. the administrative region often “blurs” underlying spatial patterns
- e.g. regions have very different sizes.
- cross boundary patterns are difficult to see
- Dense and sparse sub regions are averaged out (i.e. parks, harbours, recreational areas etc.).

Map: density maps



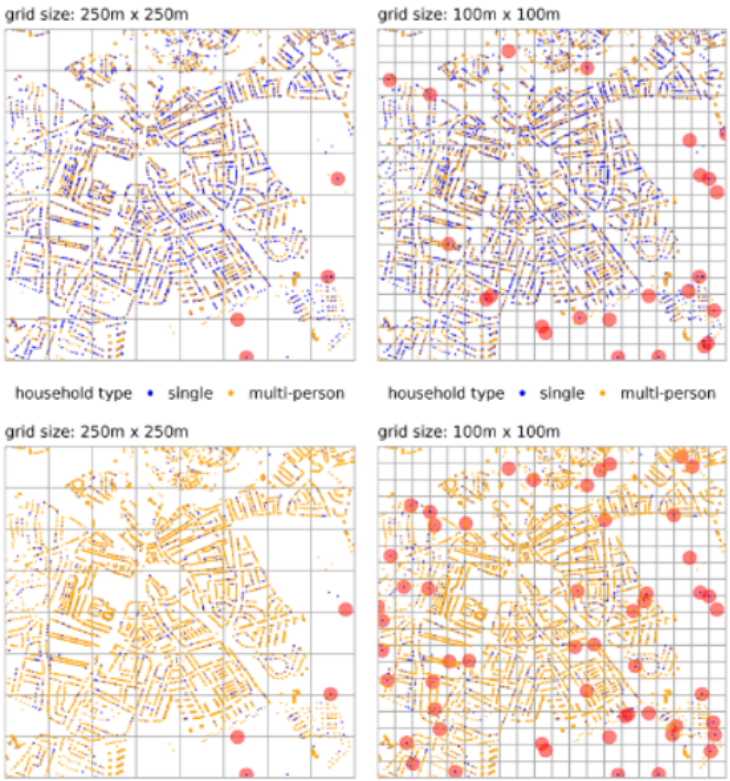
Map: density maps

- **absolute** or **relative** numbers.
- mostly univariate (multicolor)
- Uses “gridded” data, data compiled from locations, models or Earth Observation.
- Protection methods (this presentation)
- Location pertubation option, but less relevant in NL (public building register).

Demands:

- Each map has their use and should be used for that purpose.
- Density maps are under used, but have the following nice properties:
  - no administrative region size bias (distortion)
  - using locations are a base, resolution can be choosen (oops)
  - can depict absolute densities: e.g. population density, **unemployed**.
  - can depict relative densities: e.g. **unemployment rate**.
  - Both are interesting!

Resolution matters



Changing demands:

- Increasing demand for lower regional information.
- Increasing availability of location (or gridded) data (see this conference).
- So methods in this presentation focus on gridded data.

Disclosure control:

- **Risk** a measure on the severity and probability of individual disclosed. For simplicity we take here uniform severity and use probability
- **Utility** a (often not formalized) measure on useful information in the data.
- **Information loss**, the information lost after the protection. (minimize. . .)

Disclosure risk

For density maps:

Disclosure risk scenario: Plotting a variable on a map:

- The location of a population unit is very identifying.
- zooming in makes it easy to pinpoint the exact location of a population unit.
- The value of the variable at a certain location may be sensitive information.
- Densely of sparsely populated regions, implies identifiability

Disclosure scenario

Attacker:

- identifies hot / cold spots: locations that have extreme values or low density.
- Zooming in the map
- Identifies locations of individuals
- Connects the value with the individuals.

Risk measure (general)

For each generic area A (can be overlapping)(Wolf and Jonge 2017):

- the number of units n, should be at least k ((n,k)-anonymity)
- the largest value of a unit should be lower then p% of the total value (p% dominance rule).

Information loss

- Protecting privacy is removing information
- Perfectly safe data is publishing no data
- Strive for maximum utility (or information), while removing risk (removing information).
- Note: maximum utility is not necessarily original information: raw individual data can be noisy, incomplete, etc.

Information loss measures

Comparing:

- orginal grid data with protected grid data.

From (Möhler et al. 2024)

Comparing individual cell differences:

- Hellinger Distance, MSE, MAD on grid.

Comparing spatial distribution on grid:

- Earth Moving Distance / Wasserstein distance

All very sensitive to grid resolution

Map Protection Methods (grid data)

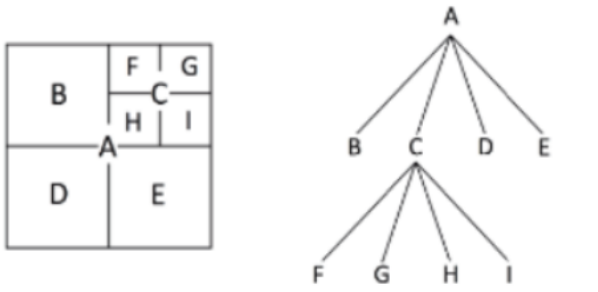
- Quadtree
- Kernel Density Smoothing
- *Multi Resolution Adaptive Smoothing*

Quadtree

Quadtree method (Lagonigro, Oller, and Martori 2017)

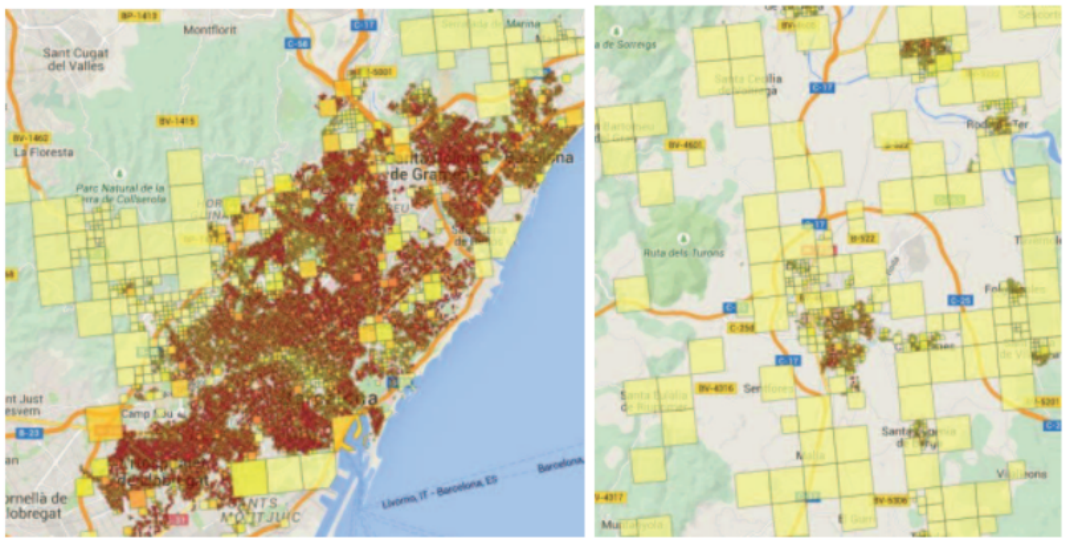
- Originally for **absolute** data (population density)
- Extended for **relative** data (de Jonge and de Wolf 2022)
- Starting area A_{ij} each grid cell in highest resolution
- If Aij not safe, aggregate 4 neighboring cells: lower resolution cell - Repeat until safe.
- Simple and safe procedure

Quadtree

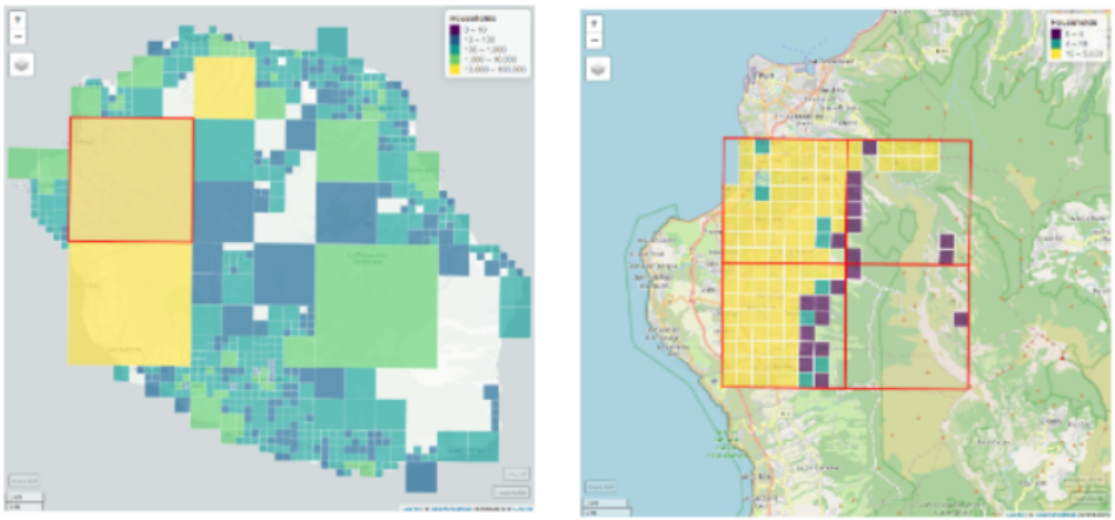


Quadtree

from (Lagonigro, Oller, and Martori 2017)



Quadtree disadvantages:



from: (Möhler et al. 2024)

Quadtree disadvantages:

- Loss of spatial detail
- Blocky result (Mondrian-like)
- Not translation invariant.

Kernel Density Smoothing

Idea:

- Use a Kernel Density Smoother to protect data (de Jonge and de Wolf 2022).
- "lend" value and density to neighboring lower density grid cells.
- Area A (weighted) support of kernel for each cell.
- KDE is well known/used for spatial patterns e.g. (Bithell 1990)

Kernel Density Smoothing

- Uses a kernel (i.e. Gaussian/Epinachov) with a band width

$$f_h(r) = \frac{1}{h^2} \sum_{i=1}^{N} \left(\frac{r - r_i}{h}\right)$$

Actually, two smoothers...

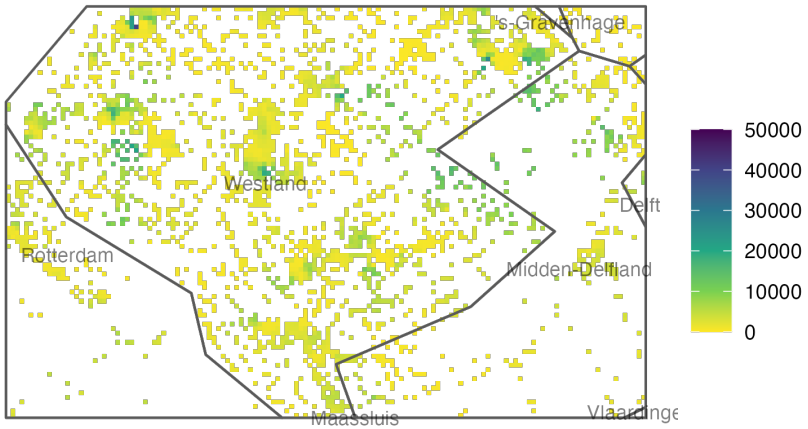
- For relative number, two smoothers:
  - variable / population
  - e.g. energy comsumption density / enterprise density

$$m_h(\vec{r}) = \frac{g_h(\vec{r})}{f_h(\vec{r})} = \frac{\sum_{i=1}^{N} a_i k\left((\vec{r} - \vec{r}_i)/h\right)}{\sum_{i=1}^{N} k\left((\vec{r} - \vec{r}_i)/h\right)}, \quad \vec{r} \in \mathcal{D}.$$

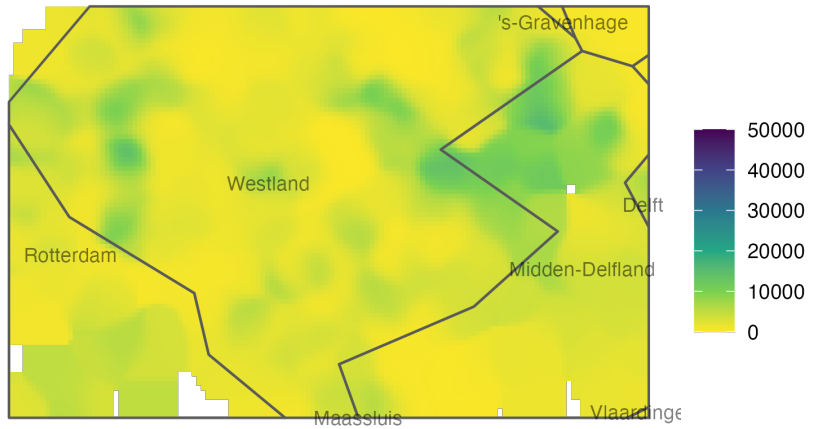
- related to spatial relative risk

KDE smoothing

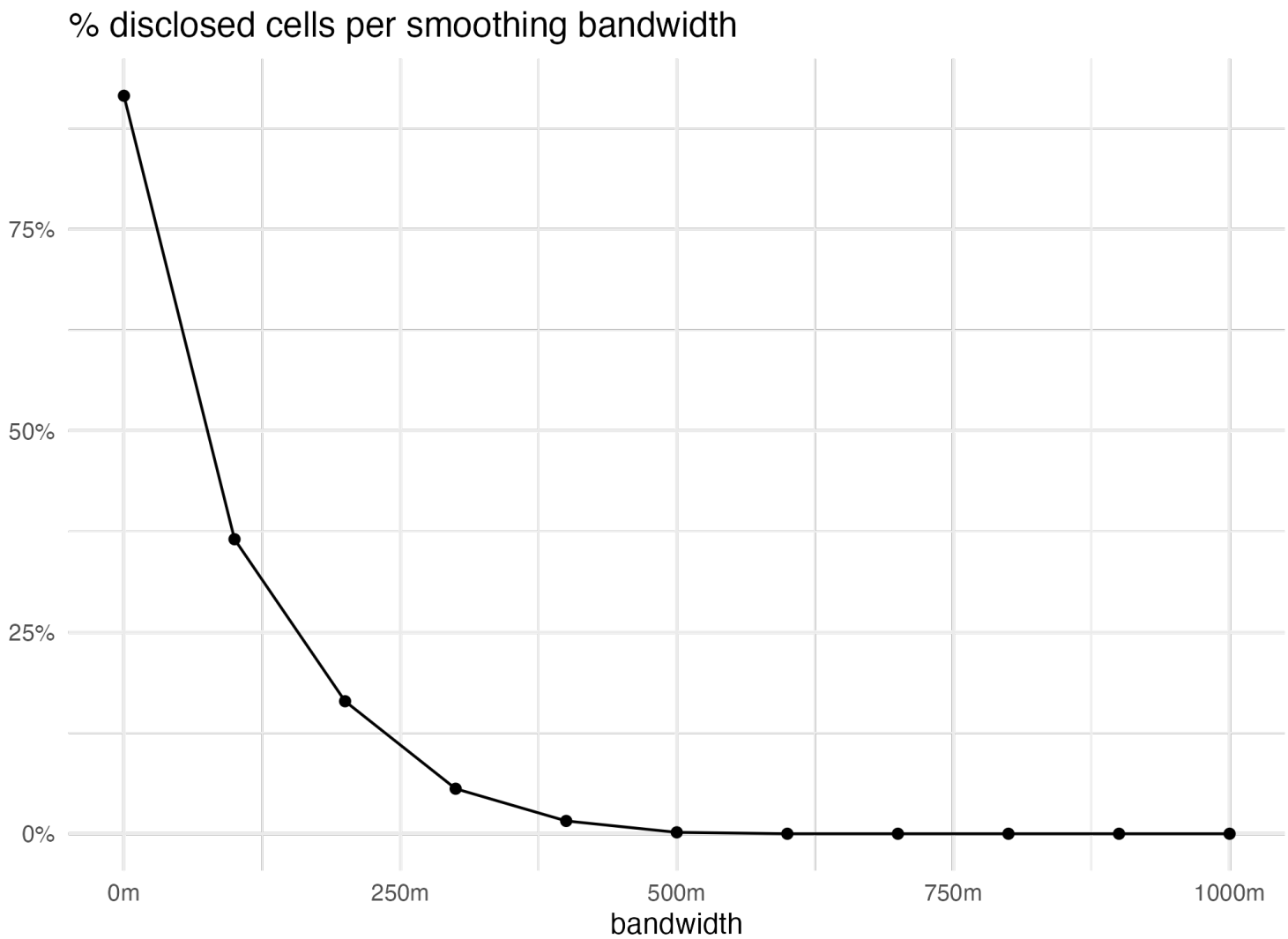
band width = 0m



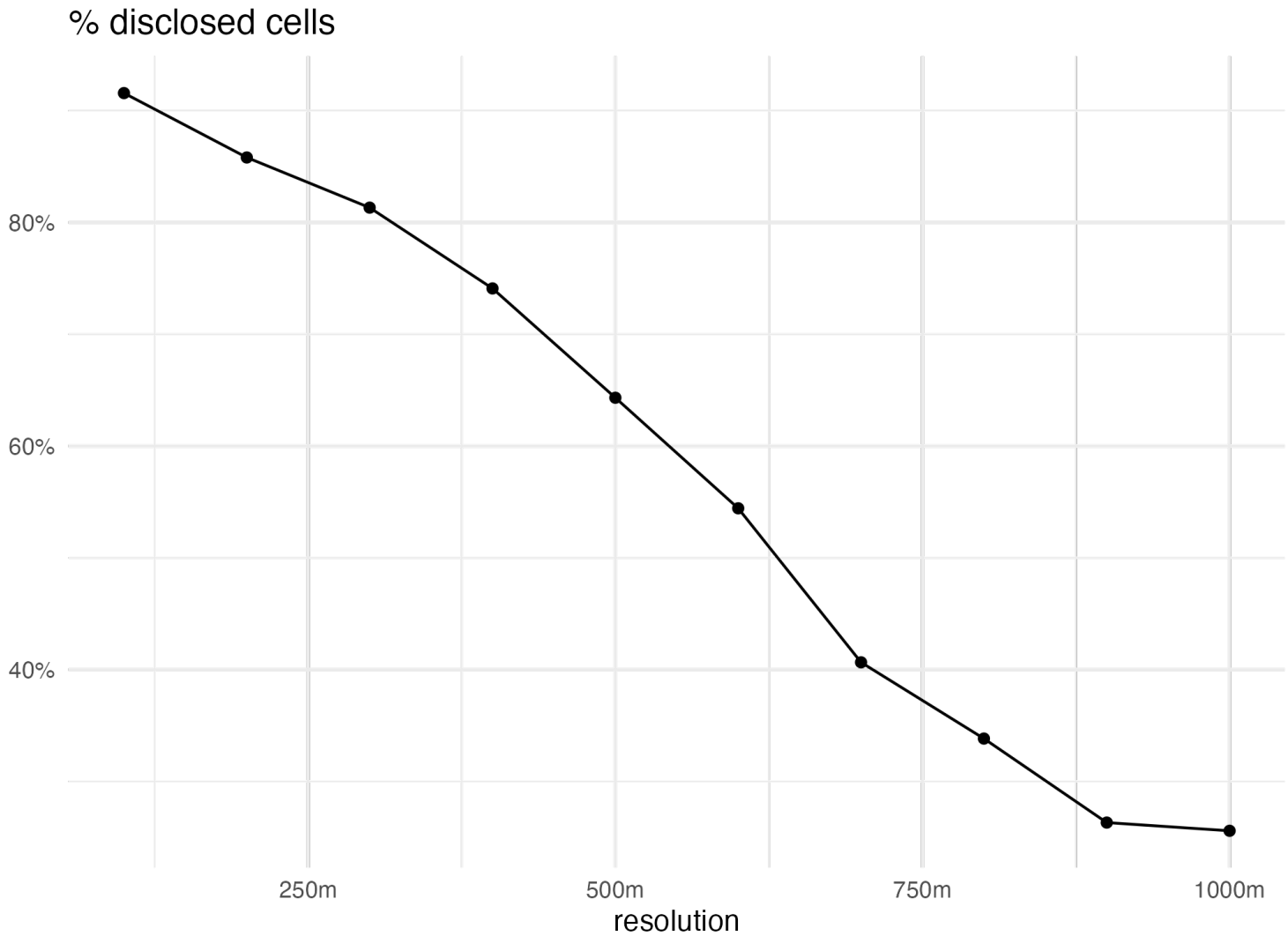
band width = 200m



KDE Dependent on band width



KDE Dependent on base resolution



Disadvantages KDE smoothing

- “No Band Width To Rule them All”. Urban and rural area’s have different scales.
- Scores bad on information loss criteria: almost every cell is adapted (more or less)
- Needs tweaking on the band width, similar to histograms / density plots.

Multiresolution Adaptive Smoothing

- Geographical processes often different scale (urban vs rural).
- KDE with fixed band width ignores that
- Quadtree misses geographical details.

Idea:

- Use multiresolution technique: e.g. wavelet
- analyse density at multiple resolutions and allow for smoothing.

Wavelets

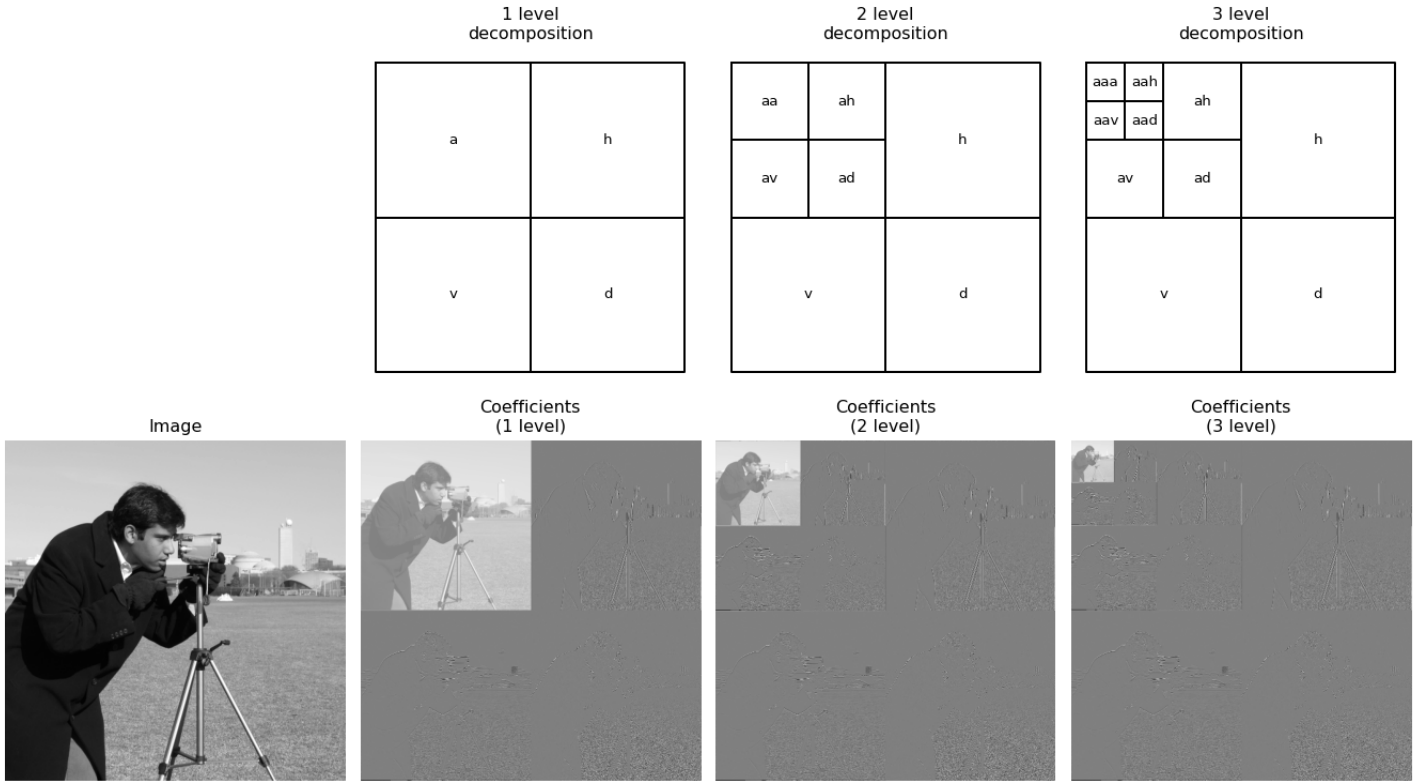
- Used for multi-resolution analysis (MRA), decompose signal / image at multiple resolutions.
- Used for denoising images: ("did I hear smoothing"?)

Also: - Used for lossy compression of images (e.g. JPEG!)

We skip the math

- Except: wavelet decomposition can have different base functions, e.g. Haar, Daubechies etc.

Wavelet and images (e.g. JPEG)



Software sdcSpatial

R package for:

- Assessing privacy problems density map
- (simple) location pertubation methods
- protect methods for density grid data.

sdcSpatial: Privacy protected maps

`sdcSpatial` and Wavelets:

- `sdcSpatial` helps to assess sensitiviy and create privacy protected density maps.
- `protect_wavelet` novel method for protecting density maps.
- takes care of rural vs urban spatial resolution.
- can be seen as combination of `protect_smooth` and `protect_quadtree`.

Sdc in `sdcSpatial`?

SDC = "Statistical Disclosure Control"

Collection of statistical methods to:

- Check if data is safe to be published
- Protect data by slightly altering (aggregated) data
  - adding noise
  - shifting mass
- Most SDC methods operate on records.
- **`sdcSpatial` works upon locations.**

Data

```
data(dwellings, package="sdcSpatial")
nrow(dwellings)
```

```
## [1] 90603
```

```
head(dwellings) # consumption/unemployed are simulated!
```

```
##        x      y consumption unemployed
## 1 149712 470104    2049.926      FALSE
## 2 149639 469906    1814.938      FALSE
## 3 149631 469888    2074.882      FALSE
## 4 149788 469831    1927.989      FALSE
## 5 149773 469834    2164.969      FALSE
## 6 149688 469898    1987.958      FALSE
```

Let's create a sdc_raster

Creation:

```
library(sdcSpatial)
unemployed <- sdc_raster( dwellings[c("x", "y")] # realistic locations
                        , dwellings$unemployed # simulated data!
                        , r = 500 # raster resolution of 500m
                        , min_count = 10 # min support
                        )
```

What has been created?

```
print(unemployed)
```

```
## logical sdc_raster object:
##    resolution: 500 500 ,  max_risk: 0.95 , min_count: 10
##    mean sensitivity score [0,1]:  0.4249471
```

42% of the data on this map is sensitive and should be protected!

Type of raster density maps:

(Stored in `unemployed$value`):

Density can be area-based:

- **number of people** per square (`$count`): population density.
- **(total) value** per square (`$sum`): number of unemployed per square.

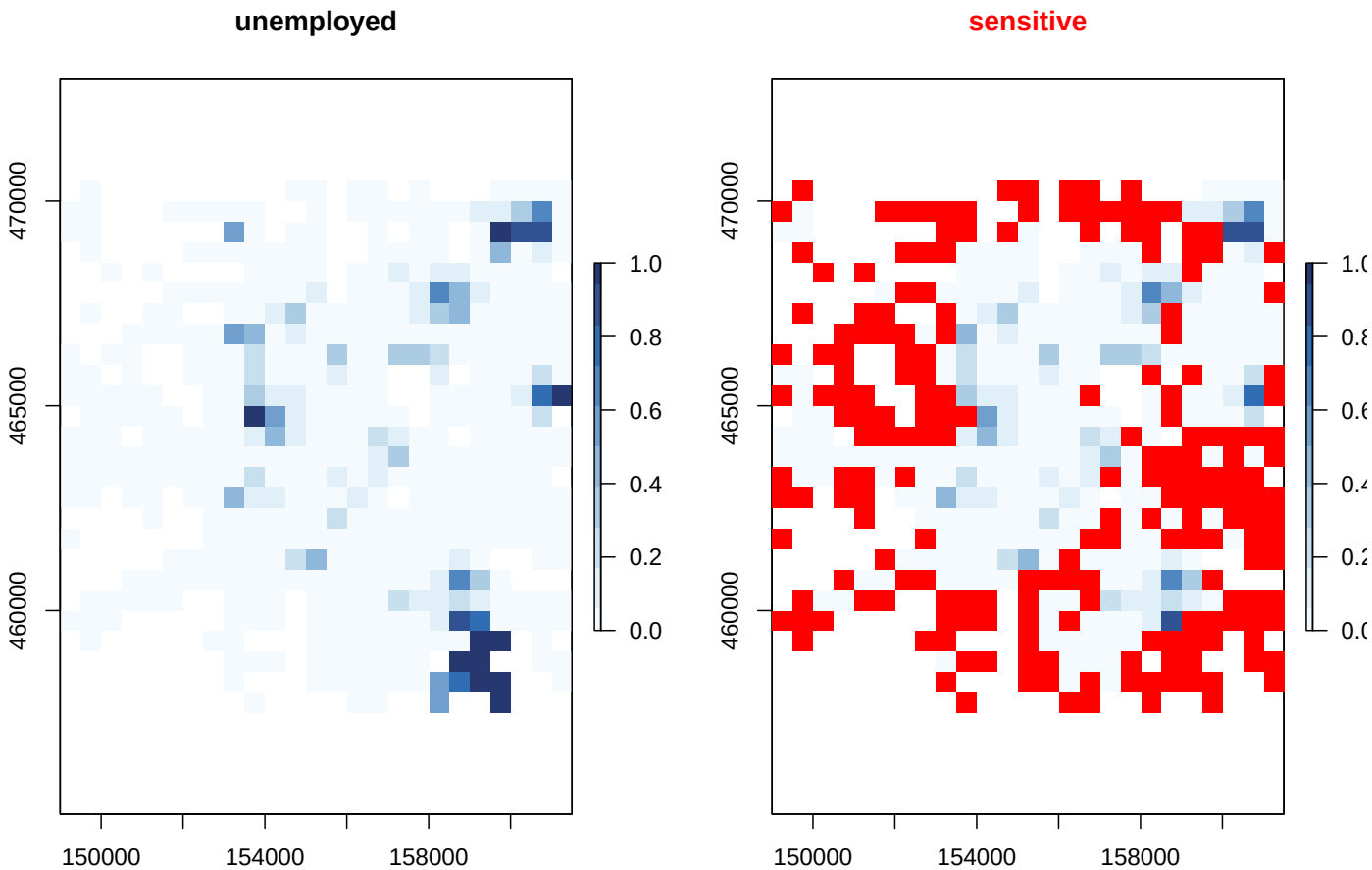
Or density can population-based:

- **Mean value** per square (`$mean`): unemployment rate per square.

Note: All density types are valid, but (total) value per square strongly interacts with population density.
(e.g. https://xkcd.com/1138).

Plotting a sdc_raster

```
plot(unemployed, "mean")
```



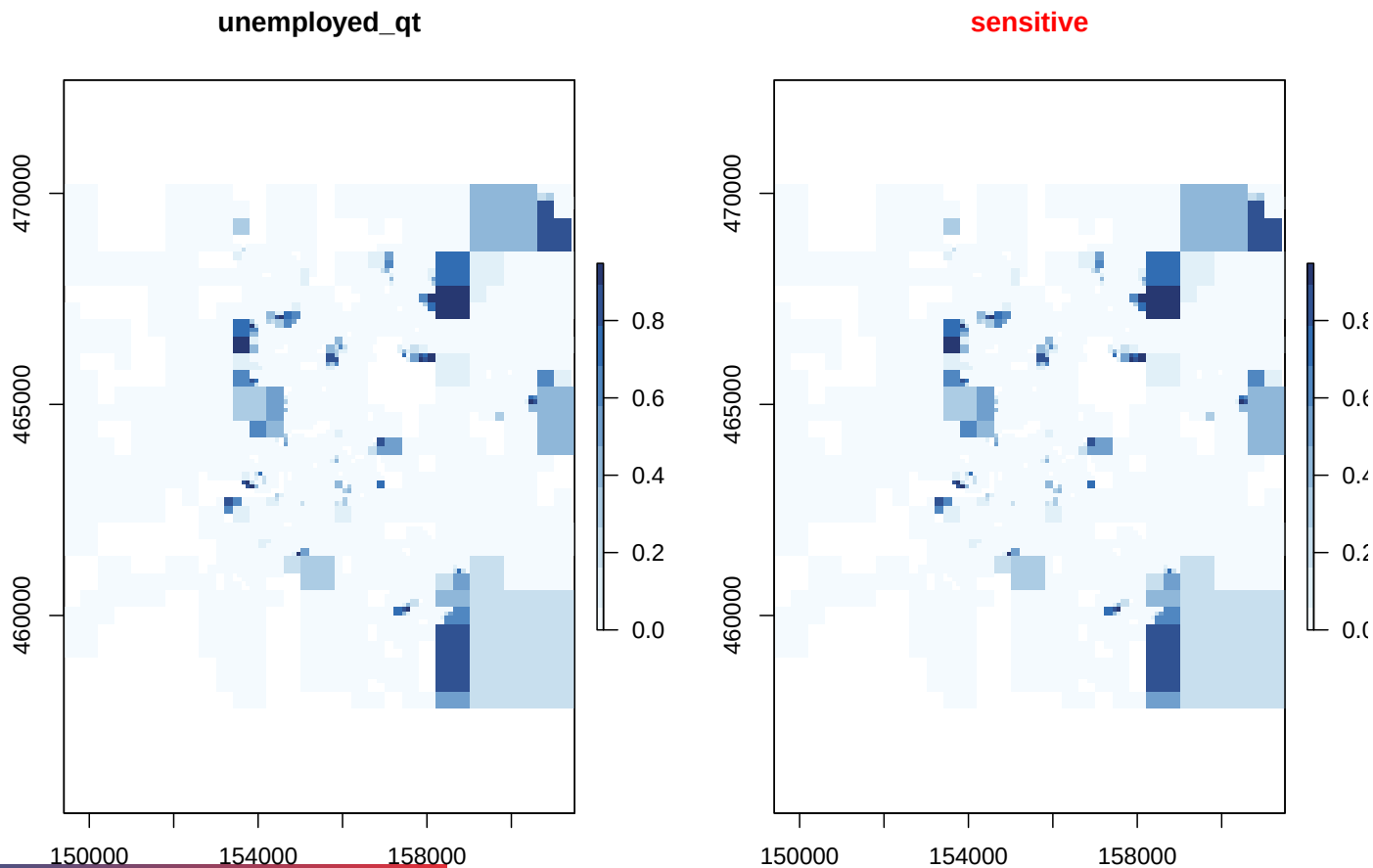
How to reduce sensitivity?

Options:

a) Remove sensitive locations: **`remove_sensitive`**.
b) Use a coarser raster: `sdc_raster`.
c) Aggregate sensitive cells hierarchically with a quad tree until not sensitive: **`protect_quadtree`** Lagonigro, Oller, and Martori (2017).
d) Apply spatial smoothing: **`protect_smooth`** (Wolf and Jonge 2018; Jonge and Wolf 2016).
e) Do a multi-resolution analysis **`protect_wavelet`**

Option: protect_quadtree

```
unemployed_100m <- sdc_raster( dwellings[c("x","y")], dwellings$unemployed
                               , r = 100) # use a finer raster
unemployed_qt <- protect_quadtree(unemployed_100m)
plot(unemployed_qt)
```



Option: `protect_quadtree`

Pro

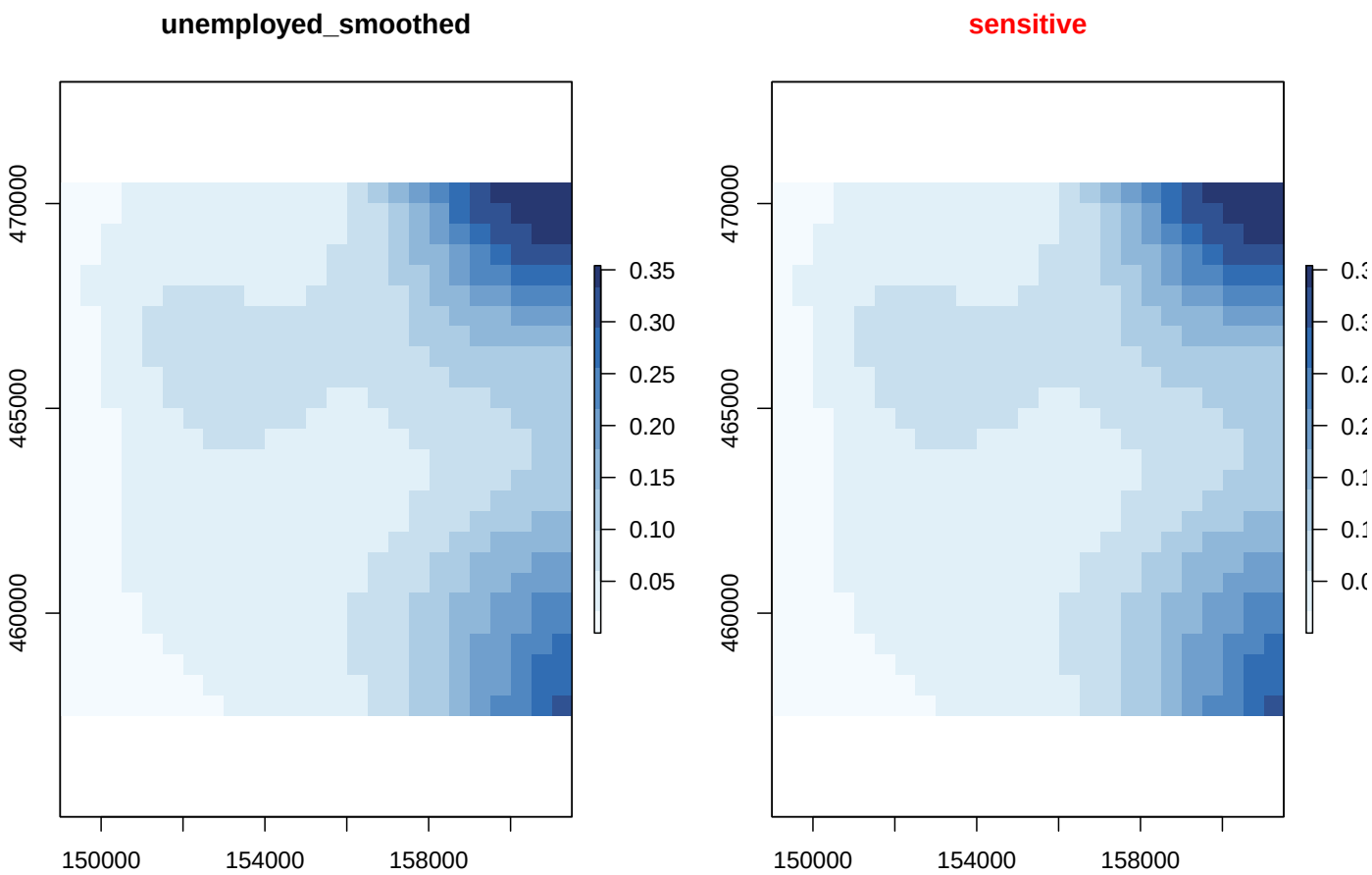
- Adapts to data density
- Adjusts until no sensitive data is left.
- It just works. . .

Cons

- Crude, it just works :-)
- Visually: “Blocky” / “Mondrian-like” / Minecraft-result
- Is not translation invariant (basis)
- Isn’t there something smoother?

Option: protect_smooth

```
unemployed_smoothed <- protect_smooth(unemployed, bw = 1500)
plot(unemployed_smoothed, "mean")
```



Option: `protect_smooth`

Pro's

- Often enhances spatial pattern visualization, removing spatial noise.
- Makes it a density map and used as source for e.g. contour map.

Con's

- Does not remove all sensitive values (depends on bandwidth `bw`)
- A fixed band width is used for all locations: may remove detailed patterns...
  spatial processes often have location dependent band widths.

Problem: smooth and adaptive

We need both a smooth and adaptive method!

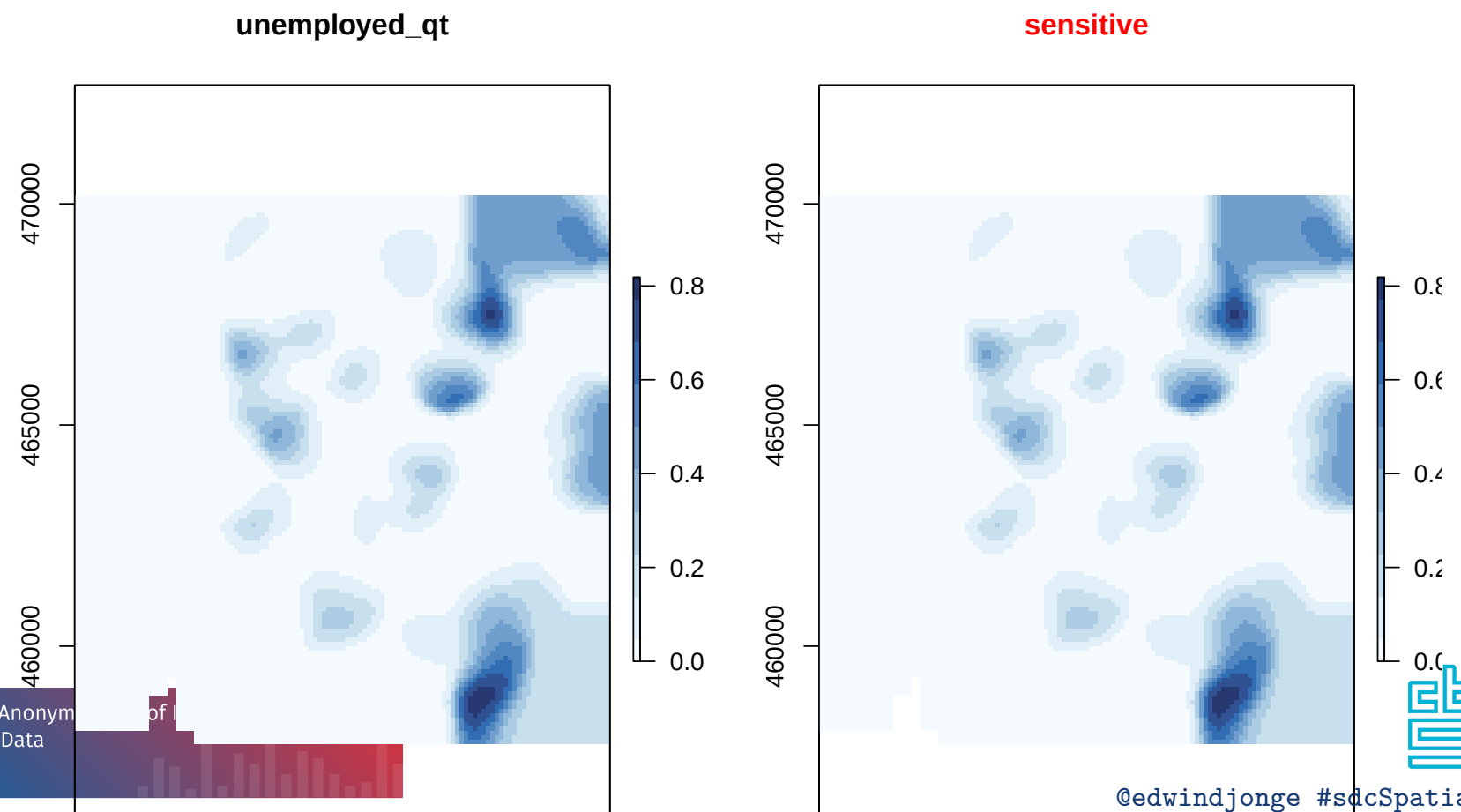
- based on multiresolution analysis

sdcSpatial and Wavelets:

- Enter: `protect_wavelet`
- Using `dwt.2d`and `waveslim::idwt.2d` of R package `waveslim` (Whitcher 2024)
- Builds a multi-resolution version of density map
- Checks sensitivity (privacy) multiple resolutions

Wavelet

```
unemployed_wvlt <- protect_wavelet(
  unemployed,
  wf = "la8", # wavelet transform / base functions
  depth = 4, # resolution depth
  ... # denoising parameters
)
```



Upcoming (December)

- `sdcSpatial` optimizations (C++, multithreaded)
  - `‘protect_smooth‘ much faster.`
  - `‘protect_quad‘ also improved performance`
- `protect_wavelet` (work in progress) on github
- “raw version”, testing make it user friendly, in December on CRAN.

The end

Thank you for your attention!

Questions?

Curious?

```
install.packages("sdcSpatial")
```

Feedback and suggestions?

https://github.com/edwindj/sdcSpatial

References

Bithell, J. F. 1990. "An Application of Density Estimation to Geographical Epidemiology." *Statistics in Medicine* 9: 691–701.

de Jonge, Edwin, and Peter-Paul de Wolf. 2022. *sdcSpatial: Statistical Disclosure Control for Spatial Data*. https://CRAN.R-project.org/package=sdcSpatial.

Jonge, Edwin de, and Peter-Paul de Wolf. 2016. "Spatial Smoothing and Statistical Disclosure Control." In *Privacy in Statistical Databases*, edited by Josep Domingo-Ferrer and Mirjana Pejić-Bach, 107–17. Springer.

Lagonigro, Raymond, Ramon Oller, and Joan Carles Martori. 2017. "A Quadtree Approach Based on European Geographic Grids: Reconciling Data Privacy and Accuracy." *SORT. Statistics and Operations Research Transactions* 41 (1): 139–58. https://doi.org/10.2436/20.8080.02.55.

Möhler, Martin, Julien Jamme, and Edwin de Jonge Peter-Paul de Wolf, Johannes Gussenbauer, and Andrej M. 2024. *Guidelines for Statistical Disclosure Control Methods Applied on Geo-Referenced Data*.

Suñé, E., C. Rovira, D. Ibáñez, and M. Farré. 2017. "Statistical Disclosure Control on Visualising Geocoded Population Data Using Quadtrees." http://nt17.pg2.at/data/x_abstracts/x_abstract_286.docx.

Whitcher, Brandon. 2024. *Waveslim: Basic Wavelet Routines for One-, Two-, and Three-Dimensional Signal Processing*. https://CRAN.R-project.org/package=waveslim.

Wolf, Peter-Paul de, and Edwin de Jonge. 2017. "Location Related Risk and Utility".