sdcSpatial: Privacy protected density maps

Multi country experiments with sdcSpatial

Edwin de Jonge / Mark van der Loo @edwindjonge

Statistics Netherlands Research & Development
@edwindjonge

uRos 2023, December 2023

sdcSpatial: Privacy protected maps

sdcSpatial: take home message

Experiments with `sdcSpatial`

- sdcSpatial Wolf and Jonge (2018),
- AT, DE, FR, NL experiments population density
- different utility measures tested
- different types of focus areas

`sdcSpatial` has methods for:

- **Creating** a **raster** map: `sdc_raster` for pop density, value density and mean density, using the excellent `raster` (Hijmans 2019).
- **Finding out** which locations are **sensitive**: `plot_sensitive`, `is_sensitive`.
- Adjusting raster map for **protecting data**: `protect_smooth`, `protect_quadtree`.
- **Removing sensitive** locations: `remove_sensitive`

uRosConf

Why sdcSpatial?

- ESS has European Code of Statistical Practice (predates GDPR, European law on Data Protection):
  no individual information may be revealed.

uRosConf

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.**

What is sdcSpatial good for?

Protecting

- *Spatial Population density*
- Spatial value density, e.g. unemployment, income
- Spatial fractions, e.g. unemployment rate
- Spatial averages, e.g. average income

We'll focus on population density

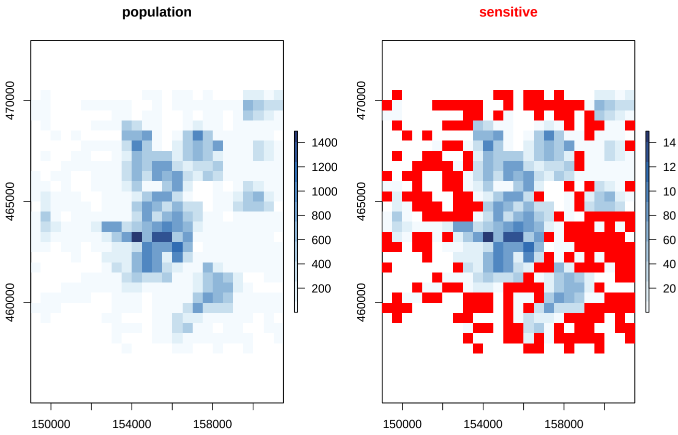
Experiments with `sdcSpatial`

Using `sdcSpatial` for AT, FR, DE and NL to protect population density, i.e. to "grid locations" with < 10 persons (Gussenbauer et al. 2023)

- 4 different area's per country: urban, moderately urban, rural, border area
- different utility measures, Hellinger distance, Moran's I
- 3 different methods: (next slides)

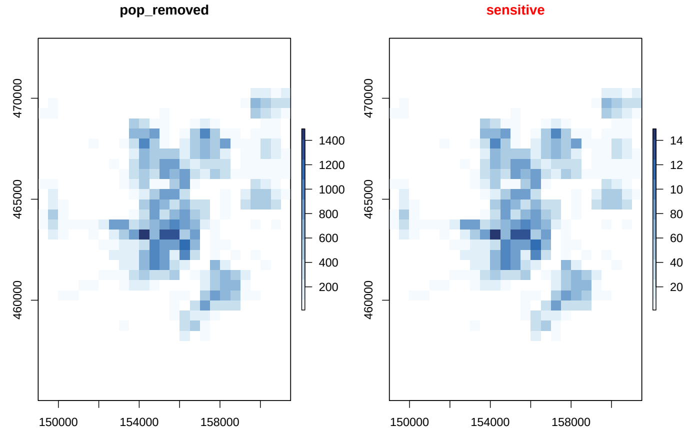
Example: unprotected

```
population <- sdc_raster( dwellings[c("x","y")]
                        , variable = 1
                        , min_count = 10
                        , r = 500)
plot(population, value="count")
```



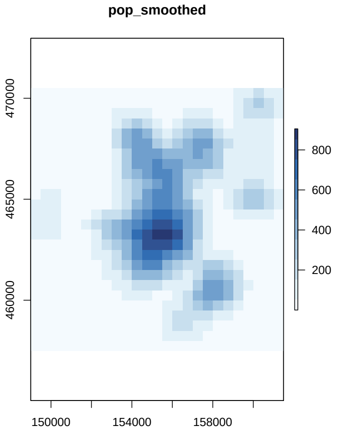
Cell removal: `remove_sensitive`

- Just remove the unsafe cells
- Pro: simple, fast, no artifacts introduced
- Con: looses mass, low density areas are removed



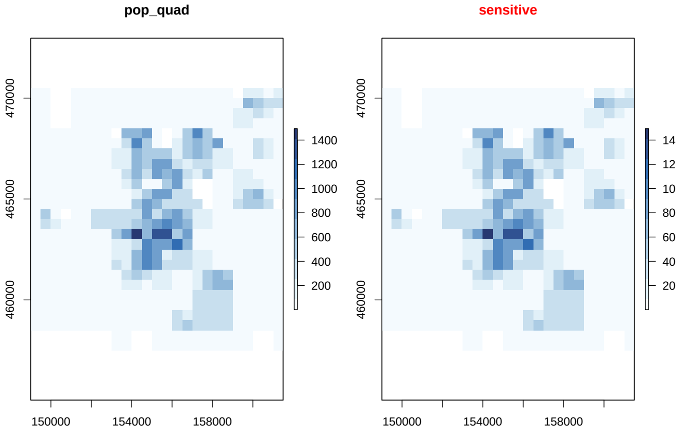
Smoothing: `protect_smooth`

- Spatial smoothing, unsafe locations are “blurred”
- Pro: highlights spatial pattern, removes spatial noise
- Con: Loose mass at edges, Introduces unplausible locations.



Quad tree: `protect_quadtree`

- Group unsafe cells, until safe
- Pro: protection garanteed, relative low adaption.
- Con: blocky result



Utility measures

Hellinger Distance (HD)

- See Map as a table:
- Difference between unprotected and protected version (rmse diff of normalized data)
- Note: does not take spatial character into account!

Kantorovic-Wasserstein Distance (KWD)

- Spatial distribution distance
- aka: earth movers distance, minimal spatial distribution adjustment
- using R package `SpatialKWD`
- Note: computational intensive

Areas countries

country	focus area	size (cells)	size (km^2)	initial risk % cells	initial risk % pop.
AT	Vienna & Suburbs	85×85	1806.25	10.09	0.03
	Bregenz	39×39	380.25	9.57	0.08
	Alps in Tyrol	73×73	1332.25	17.10	0.59
	Krems an der Donau	41×41	420.25	12.38	0.28
DE	Ruhr valley	55×55	756.25	3.9	0.02
	Mainz & Wiesbaden	41×41	420.25	9.2	0.04
	Strelasund region	75×75	1406.25	22.2	0.64
	German Allgäu	55×55	756.25	24.4	1.06
FR	Saint-Denis	45×45	81.00	31.4	2.43
	Saint-Pierre	109×109	475.24	49.1	8.63
	La Plaine	41×41	67.24	72.2	31.63
	Saint-Gilles	71×71	201.64	51.0	10.31
NL	Amsterdam	59×46	678.50	12.1	0.04
	Almere	47×42	493.50	13.9	0.07
	Drenthe	89×111	2469.75	21.7	0.64
	Parkstad	31×44	341.50	11.1	0.09

Results (NL)

focus area (NL)	method	residual risk		utility	
		% cells	% pop.	HD	KWD
Amsterdam	removal	0	0	.01	.004
	quad tree I	8.1	0.01	.08	.015
	quad tree II	0.8	< .01	.13	.054
	smoothing	1.1	< .01	.22	.257
Almere	removal	0	0	.02	.009
	quad tree I	15.9	0.03	.09	.018
	quad tree II	1.8	< .01	.13	.054
	smoothing	1.3	< .01	.25	.316
Drenthe	removal	0	0	.06	.080
	quad tree I	13.2	.13	.16	.062
	quad tree II	0.3	< .01	.23	.164
	smoothing	0.6	< .01	.31	.407
Parkstad	removal	0	0	.02	.007
	quad tree I	6.6	0.01	.13	.039
	quad tree II	0	0	.20	.124
	smoothing	0	< .01	.27	.352

Discussion

- Using HD sdc-table-like measures, cell removal best, smoothing worse
- Same with spatial KWD!

However. . .

- current utility measure look at individual differences
- It does not take spatial pattern/shape into account
- Smoothing may *create* utility, reducing noise.
- So looking into other utility spatial utility measures.

Upcoming changes sdcSpatial 0.7

- Hellinger Distance (HD): `distance_hellinger()`
- speed improvements:
  - raster construction (faster)
  - smoothing (much more efficient): from city size to country size.
- tbp in January 2024

uRosConf

Thank you for your attention!

Questions?

Curious?

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

Feedback and suggestions?

https://github.com/edwindj/sdcSpatial/issues

References

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

Gussenbauer, Johannes, Julien Jamme, Edwin de Jonge, Peter-Paul de Wolf, and Martin Mohler. 2023. "Spatial SDC Experiments and Evaluations with Multiple Countries Comparison." Presented at UNECE/Eurostat worksession Statistical Data Confidentiality, 26–28 September, Wiesbaden. https://unece.org/sites/default/files/2023-08/SDC2023_S3_4_Austria_Gussenbauer_D.pdf.

Hijmans, Robert J. 2019. *Raster: Geographic Data Analysis and Modeling*. https://CRAN.R-project.org/package=raster.

Wolf, Peter-Paul de, and Edwin de Jonge. 2018. "Spatial Smoothing and Statistical Disclosure Control." In *Privacy in Statistical Databases - PSD 2018*, edited by Josep Domingo-Ferrer and Francisco Montes Suay. Springer.