Kernel density analysis – a tool for the visualization of spatial patterns in regional studies

Robust-Workshop in Essen, 21st February, 2019

Rüdiger Budde

Theoretical background

Kernel Density Estimation (KDE)

Kernel Density Estimation (KDE) is a non-parametric technique for density estimation in which a known density function (the kernel) is averaged across the observed data points to create a smooth approximation.

Density Estimation and Histograms

Let *b* denote the bin-width then the histogram estimation at a point x from a random sample of size *n* is given by,

$$\hat{f}_H(x;b) = \frac{number\ of\ observations\ in\ bin\ containing\ x}{nb}$$

Two choices have to be made when constructing a histogram:

- Positioning of the bin edges
- Bin-width

KDE – Smoothing the Histogram

Let $X_1, \ldots, X_n$ be a random sample taken from a continuous, univariate density *f*. The kernel density estimator is given by,

$$\hat{f}(x;h) = \frac{1}{nh}\sum_{i=1}^{n} K\{(x - X_i)/h\}$$

- *K* is a function satisfying $\int K(x)\,dx = 1$
- The function *K* is referred to as the *kernel*.
- *h* is a positive number, usually called the *bandwidth* or *window width*.

Kernels

- Gaussian
- Epanechnikov
- Rectangular
- Triangular
- Biweight
- Uniform
- Cosine

Refer to Table 2.1 Wand and Jones, page 31.

… most unimodal densities perform about the same as each other when used as a kernel.

- Use the **Gaussian** kernel.

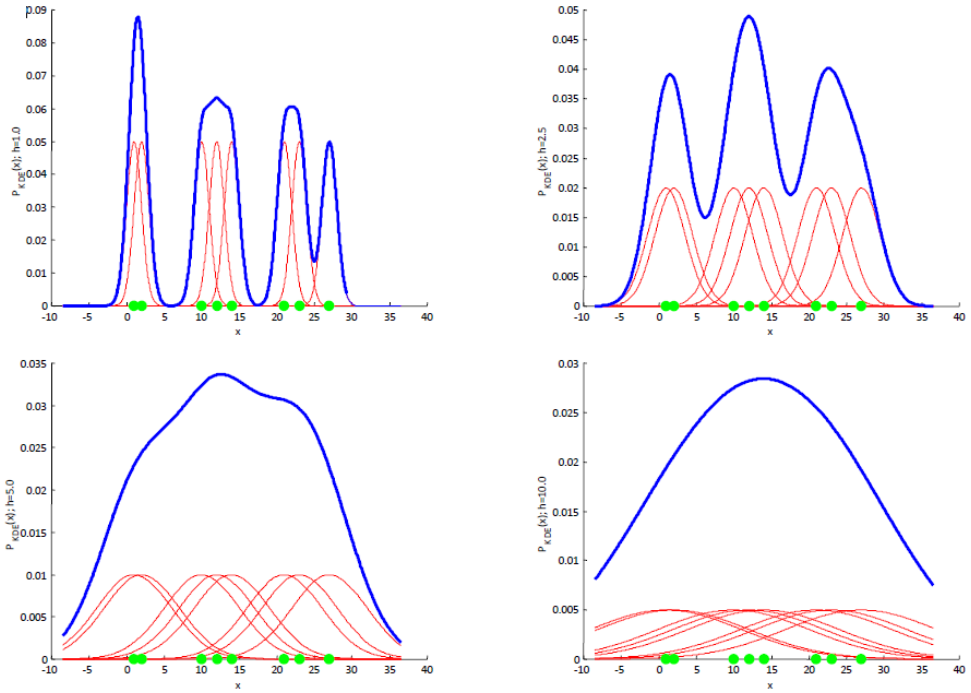
Wand M.P. and M.C. Jones (1995), *Kernel Smoothing,* Monographs on Statistics and Applied Probability 60, Chapman and Hall/CRC, 212 pp.

Bandwidth selection

The problem of choosing $\boldsymbol{h}$ is crucial in density estimation

–A large h will over-smooth the DE and mask the structure of the data

–A small h will yield a DE that is spiky and very hard to interpret



Pattern Analysis | Ricardo Gutierrez-Osuna | CSE@TAMU

Optimal bandwith

Theoretical Constructs

- Mean Integrated Squared Error (MISE)
- Asymptotic Mean Integrated Squared Error (AMISE)The formulas are not able to be used directly since they involve the unknown density function.

A rule-of-thumb bandwidth estimator

- If Gaussian basis functions are used to approximate univariate data, and the underlying density being estimated is Gaussian, the optimal choice for h

$$h = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{\frac{1}{5}} \approx 1.06\hat{\sigma}n^{-1/5},$$

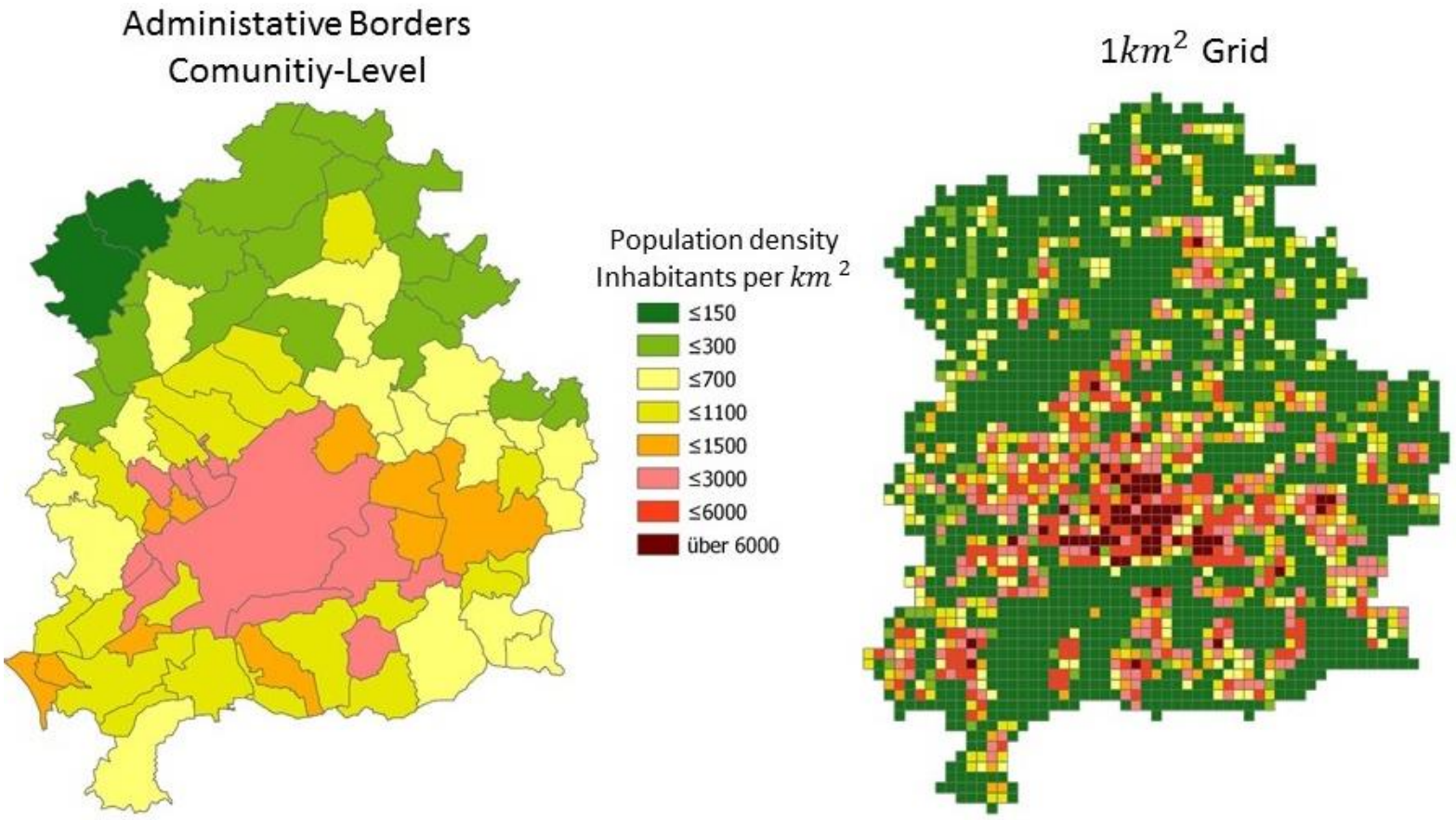
σ̂ standard deviation of the samples.

A rule-of-thumb bandwidth estimator. where is the standard deviation of the samples. The estimate based on the rule-of-thumb bandwidth is significantly oversmoothed.

Generally these test are constructed for samples with 1 dimension, while spatial problems have got at least two dimensions.

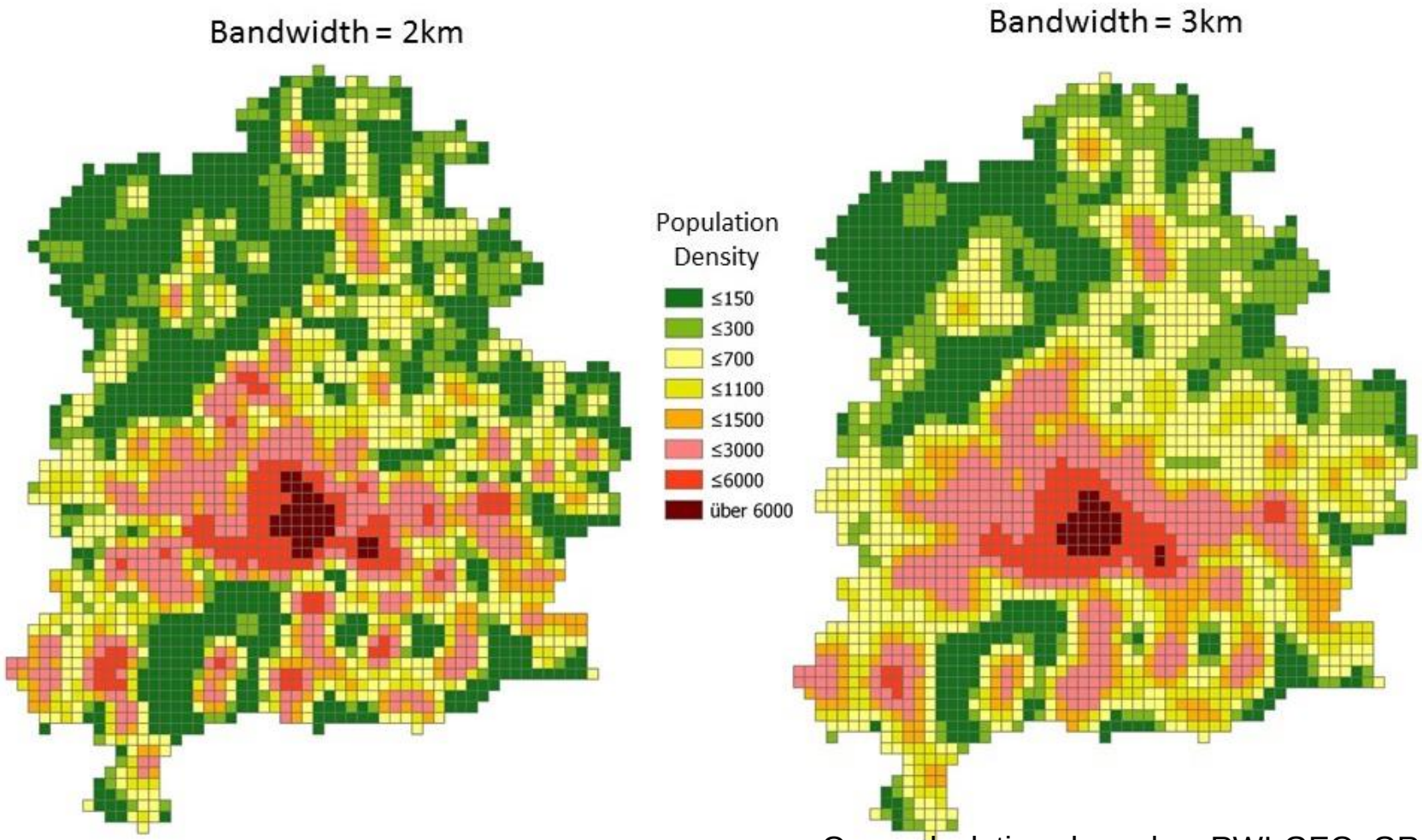
Jukka M. krisp, stefan peters, Christian e. Murphy & hongChao fan, Visual Bandwidth Selection for Kernel Density Maps, Geoinformation 5/2009, S. 445–454

Practical application of Kernel Density Analysis

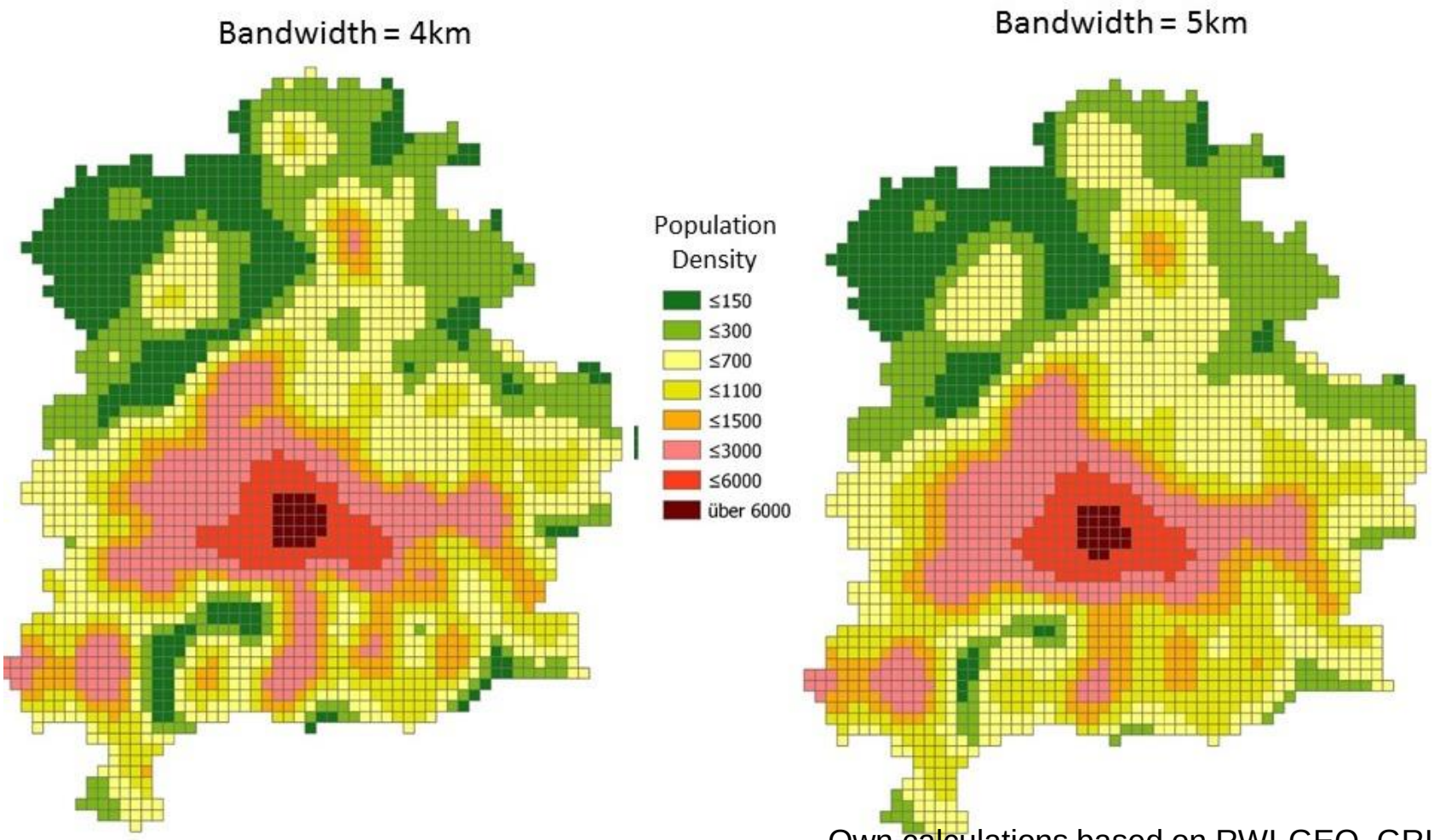


Own calculations based on © GeoBasis-DE / BKG 2013 or RWI-GEO_GRID.

Own calculations based on RWI-GEO_GRID.

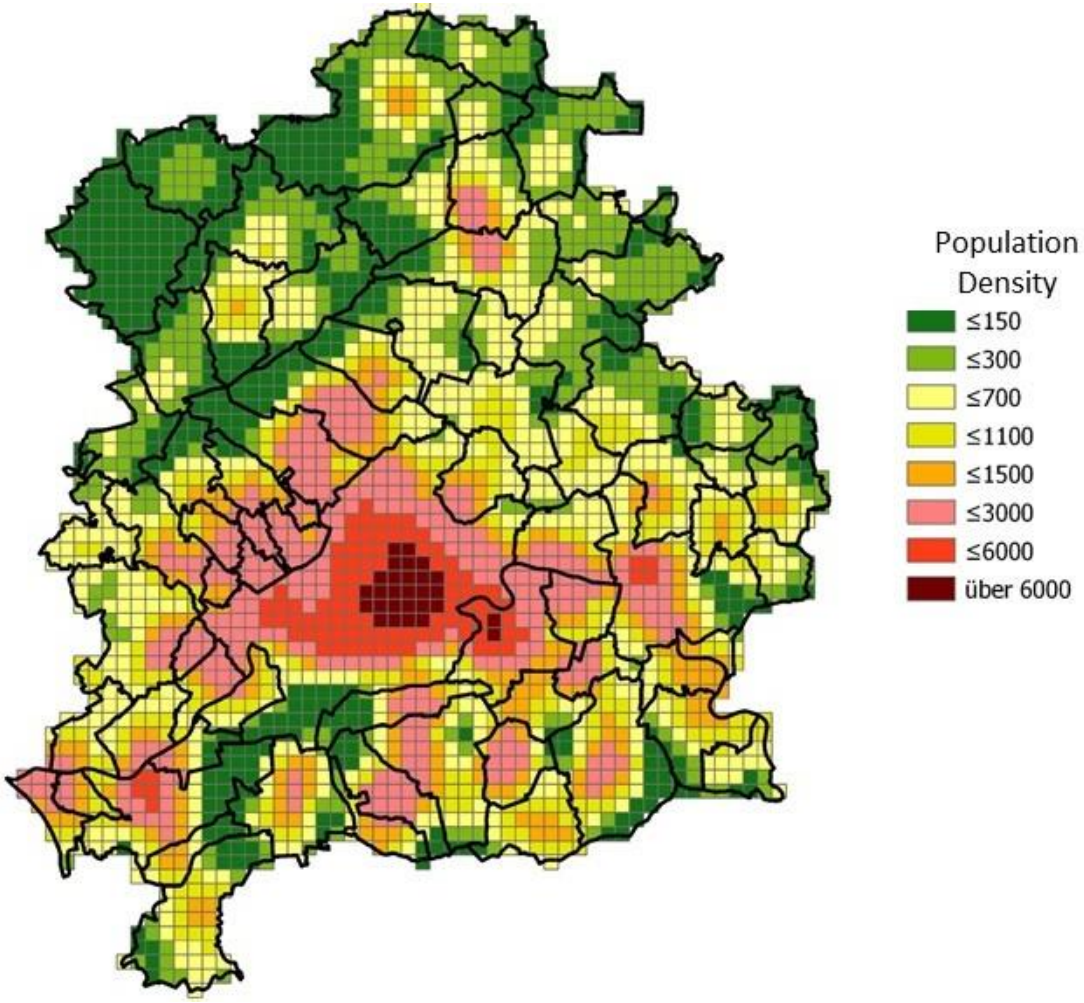


Own calculations based on RWI-GEO_GRID.



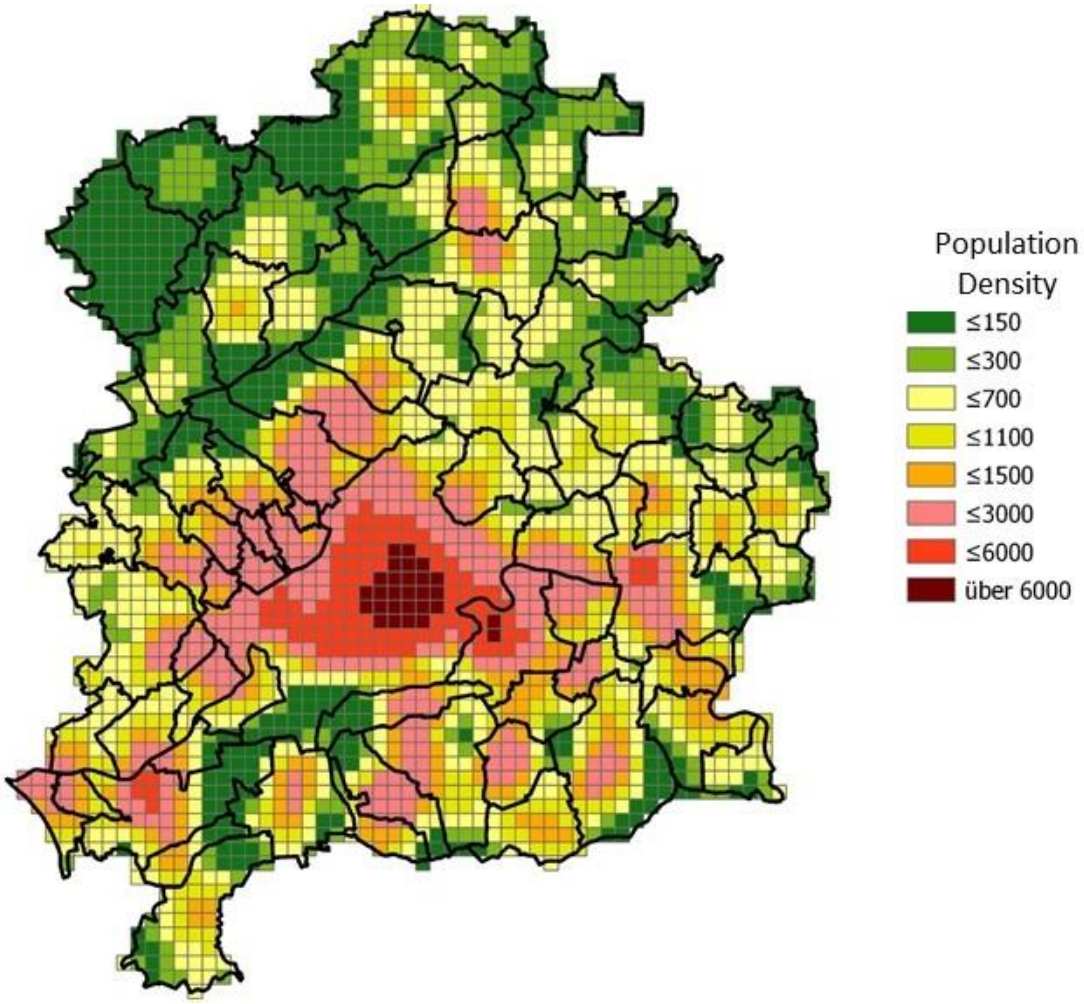
Own calculations based on RWI-GEO_GRID.

Original Distribution		Bandwidth			
		2	3	4	5
Population density	Number of spatial units	Number of spatial units			
Up to 150	54	52	29	19	8
150 up to 300	121	165	87	39	20
300 up to 700	163	159	78	36	18
700 up to 1100	116	144	90	42	13
1100 up to 1500	107	110	84	41	19
1500 up to 3000	155	59	17	6	4
3000 up to 6000	101	19	3	1	1
6000 and more	23	2	2	1	1
total	840	710	390	185	84
absolute reduction of spatial units		-130	-320	-205	-101



Own calculations based on RWI-GEO_GRID.

Application of the found results



Own calculations based on RWI-GEO_GRID.

Analytic regions versus administrative regions

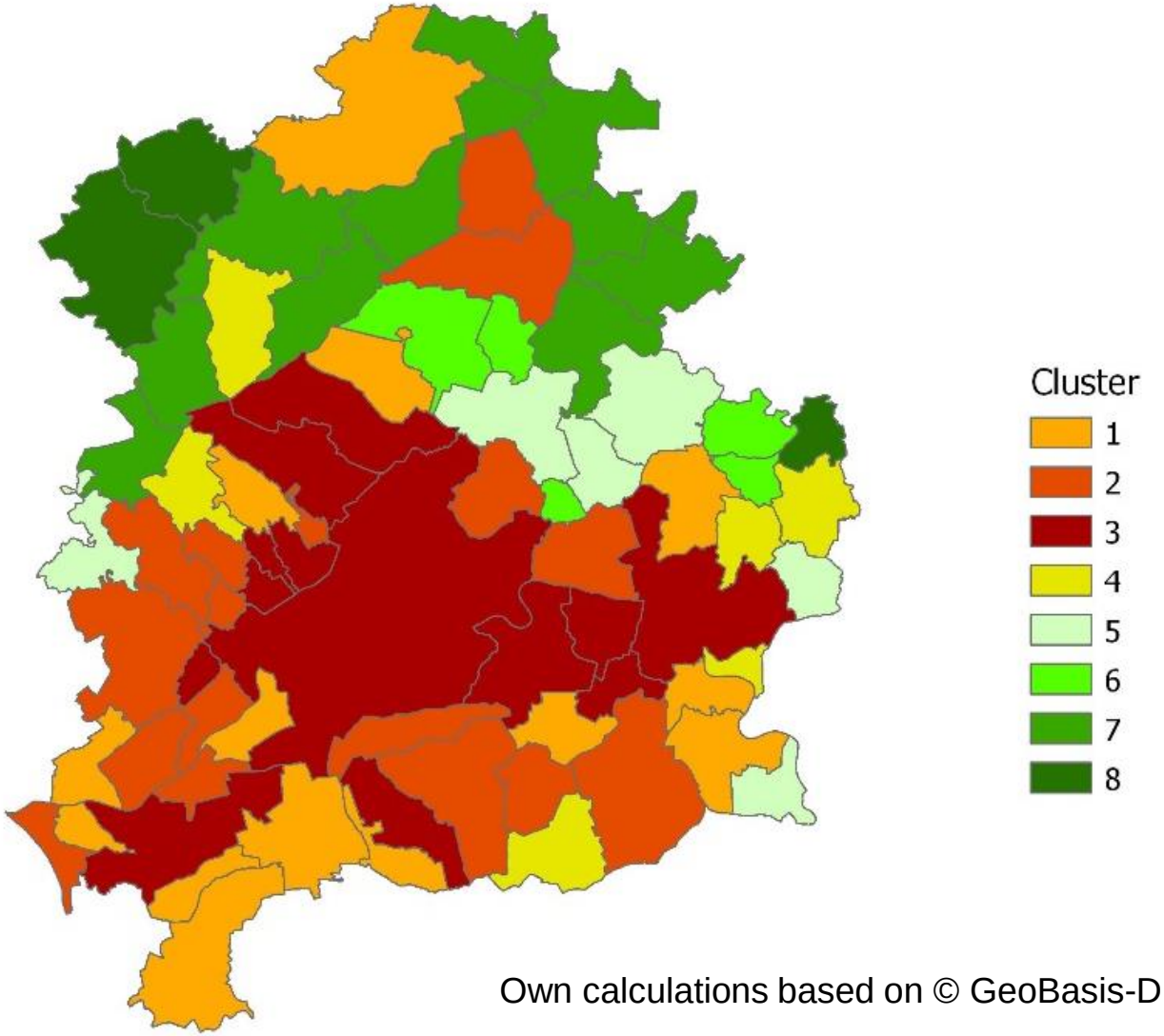
- For each municipality, the proportions of the total area can be determined in the density classes.
- In order to further compress the information and to summarize similarly structured municipalities, the municipalities have been clustered according to these shares.
- For this purpose it seems reasonable to consider 8 clusters. They are characterized in the following table by the mean values of the individual parts.

Population density in changed class sizes

Category	Population density, estimated by KDE
Rural Areas	<=300
Peri-Urban 1	300 < x<= 700
Peri-Urban 2	700 < x<= 1100
Peri-Urban 3	1100 < x<= 1500
Metropolitan Areas	>1500

Regional classification after clustering

Cluster	Number of communities	Rural Areas	Peri-Urban 1	Peri-Urban 2	Peri-Urban 3	Met
1	14	2,5	7,4	17,3	61,0	11,8
2	17	1,0	8,0	9,8	21,4	59,7
3	13	0,0	1,8	3,1	5,0	90,2
4	6	0,4	9,0	49,7	40,7	0,2
5	7	6,1	46,9	47,0	0,0	0,0
6	5	4,1	95,9	0,0	0,0	0,0
7	11	46,9	51,1	2,0	0,0	0,0
8	4	100,0	0	0,0	0,0	0,0



Own calculations based on © GeoBasis-DE / BKG.

Thank you, for your attention!