Simulated geo-coordinates as a general means for regional analysis: theory and examples

Ulrich Rendtel (FU Berlin)
Timo Schmid (FU Berlin)
Marcus Gross (INWT-Statistics Berlin)
Kerstin Erfurth (Amt für Statistik Berlin/Brandenburg)
Nikos Tzavidis (Univ. Southampton)

6. June 2019
Italian Conference on Survey Methodology (ITACOSM 2019)
University of Florence

Freie Universität Berlin

Local aggregates: becoming a frequent data type

- Polling results at the level of voting districts
- Aggregates for ZIP-Code areas at different level due to confidentially reasons
- Open data at very low regional level as background data, for example, demographic data for urban planning districts.
- Inspire data on gridsof different size (100m $\times$ 100m,...,1km $\times$ 1km,...)

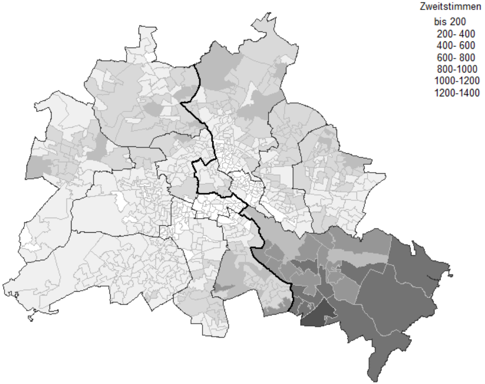
The display of regional concentrations with local aggregate data

With Choropleths

- Homogeneous distribution within local areas is unrealistic
- Jumps at area borders are unrealistic
- Discrete display of levels, usually 5 different colours, hides information.
- Representation of local units by their area is unrealistic, especially for polling districts. All polling districts represent the same number of voters but are of quite different size.
- No obvious display of regional concentrations with local aggregates.

The display of regional concentrations with local aggregates data: a Choropleth representation of voting results

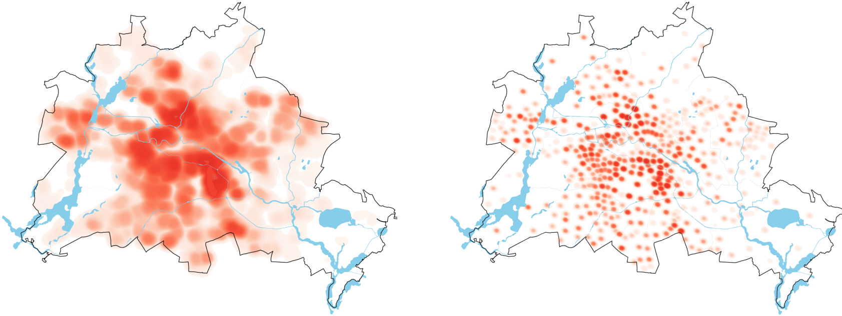
Is the South-East of Berlin a stronghold of AfD voters, an emerging right wing party?



The display of regional concentrations with local aggregate data: using center of area as geo-coordinate

With kernel density estimates

- Smooth shape.
- Regional concentration regions can be identified by highest density regions.
- The shape depends heavily on the smoothing parameter. However, much theoretical and computational support (Wand/Jones 1994)
- However, the exact geo-coordinates have to be known!
- The simple strategy to concentrate units in the center of the area fails.



Kernel density estimates of the population with migration background in Berlin. **Left:** with exact geo-coordinates **Right:** with aggregates assumed at the center of local planning units (LORs)

A Nonparametric Approach

$\mathbf{W} = \{W_1, \ldots, W_n\}$ coarse measurement of exact geo-coordinates
$\mathbf{X} = \{X_1, \ldots, X_n\}$
$\pi(W|X) = \prod_{i=1}^{n} \pi(W_i|X_i)$, with:

$$\pi(W_i|X_i) = \begin{cases} 1 & \text{for } X_i \in \mathit{area}(W_i) \\ 0 & \text{else.} \end{cases} \tag{1}$$

Pseudo samples (simulated geocoordinates) from:

$$\pi(X_i|W_i) \propto \pi(W_i|X_i)\pi(X_i). \tag{2}$$

Alternative literature on "Change of Support": parametric approach (Poisson for counts) with constant intensity function over areas: Bradley et al. (2016) in JASA

The simulation of geo-coordinates (1/2)

Basic ideas of algorithm:

1. Replace the uniform density over the area by the current density of an iterative procedure.
2. Draw a stratified sample of size equal aggregate size for each area on a fine grid. Select grid points with probability proportional to current density estimate.
3. Update kernel density from the simulated geo-coordinates.
4. Ignore the first B (=Burn-in) iterations and take the mean of the last M iterations.

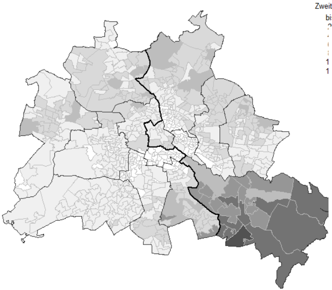
Details in Gross et al. (2017): *Estimating the density of ethnic minorities and aged people: Multivariate Kernel density estimation applied to sensitive georeferenced administrative data protected via measurement error.* Journal Royal Statistical Society (Series A), 180, 161–183

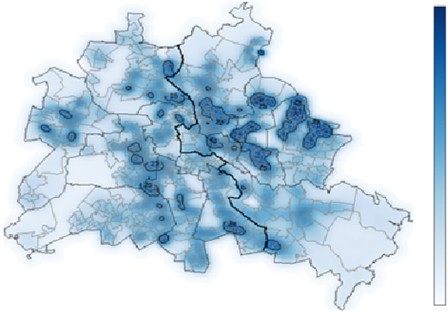
R-Package: *kernelheaping*

The simulation of geo-coordinates (2/2)

Modifications:

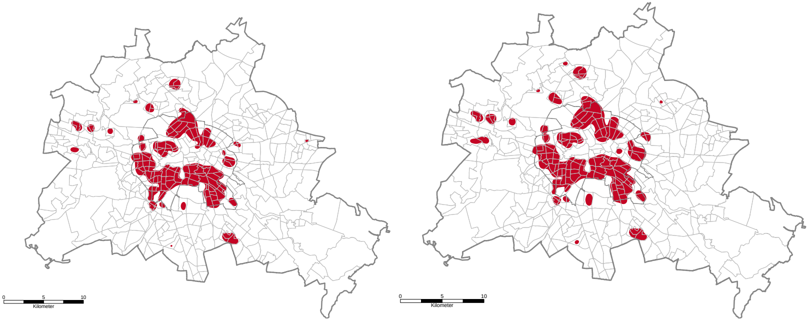
- Unsettled areas: Skip the grid points of unsettled areas (lakes, forest, parks, industry) in the algorithm!
- Boundary correction: Kernel function does cover unsettled or out-of-area regions. Rescale Kernel function to sum up to 1 over the valid grid points. Computer intensive!





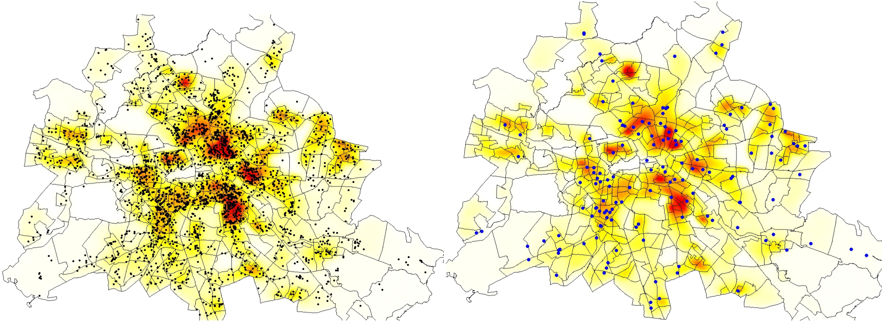
Regional concentration of AfD-Voters in Berlin elections (2016). **Left:** Traditional Choropleth map **Right:** Map constructed on the basis of Kernelheaping algorithm. Highest density regions included in graph.

Application II: Regional concentration of ethnic minorities by highest density areas



Comparison of the location of 25 % highest density regions for the population with migration background **Left:** Regions in 2007 **Right:** Regions in 2015.

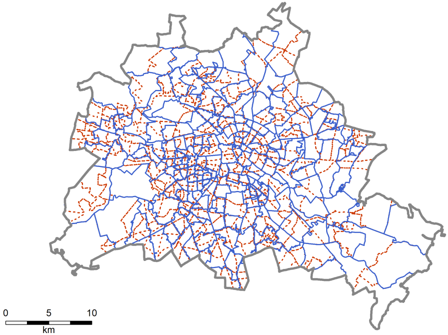
Application III: Service maps created from Open Data: Regional offer of Kindergardens and pediatrists in Berlin



Left: Distribution of children (age ≤ 6 years) and the location of Kindergardens **Right:** Distribution of children (age ≤ 18 years) and the location of pediatrists.

Application IV: Disaggregation in non-hierachical administrative area systems (1/3)

- System at start: 193 ZIP codes (Blue lines)
- Target system: 447 administrative planning areas (LORs) (Red lines)



Task: Redistribute student resident numbers at ZIP-Level (from enrollment office) to administrative planning areas (LORs)!

Application IV: Disaggregation in non-hierachical administrative area systems (2/3)

Problem can be solved with Kernelheaping Algorithm:

1. For each iteration allocate the simulated geo-coordinates to the LORs they fall into.
2. Take the mean of the case numbers over the iterations of the algorithm.

Standard disaggregation uses uniform density for re-allocation.

Application IV: Disaggregation in non-hierachical administrative area systems (3/3)

A simulation experiment for the evaluation of the Kernelheaping Algorithm:

1. Generate 250 geo-coordinates per LOR
2. Redistribute geo-coordinates to ZIP areas
3. Take ZIP totals to start the disaggregation:
    - Use Kernelheaping
    - Use uniform density allocation as baseline
4. Compute absolute percentage deviance (APD) and RMSE over 100 replications.

Method	Average APD	Average RMSE
Kernelheaping	9.8 %	33.5
Baseline	14.4%	60.6

Application V: Computation of local shares of political parties in polling

- The standard approach for reporting shares:

$$\frac{\text{no. of voters for party P in area}}{\text{no. of voters in area}}$$

- Alternative approach via kernel density estimates:

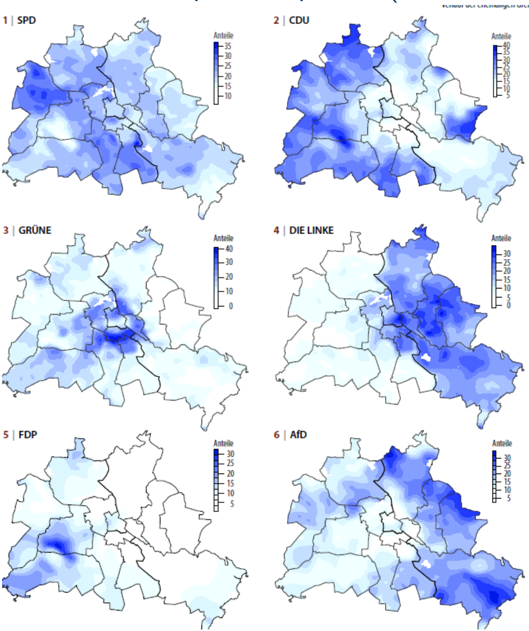
$$\frac{N_P f_P(x)}{N_V f_V(x)} \quad \text{(Nadaraya-Watson Estimator)}$$

where:

N_P	Total number of voters for party P
N_V	Total number of voters
$f_P(x)$	Kernel density estimate of voters of Party P at x
$f_V(x)$	Kernel density estimate of voters at x

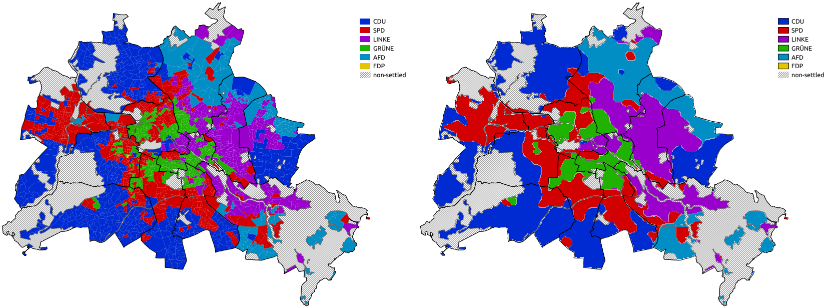
- Estimate $f_V(x)$ by the Kernelheaping Algorithm.
- In order to avoid inconsistent results the estimation of $f_P(x)$ has to be concentrated on the same geo-coordinates which were used for the computation of $f_v(x)$.

Local shares of political parties (Berlin election 2016)



Application VI: Local winners in elections

- Compare the local shares of political parties.
- The winner at coordinate x is the party with the highest local share.

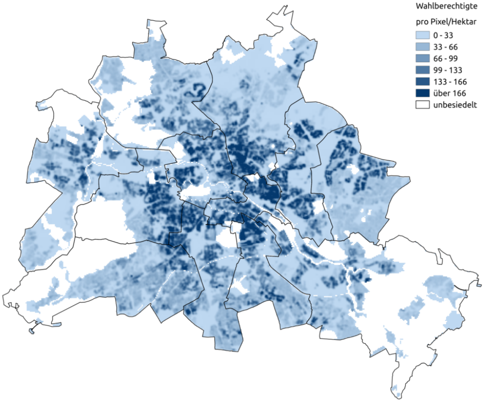


The local winner of the Federal elections 2017 in Berlin **Left:** Winner according to shares in voting districts **Right:** Winner according to local share estimation.

Evaluation: Issues

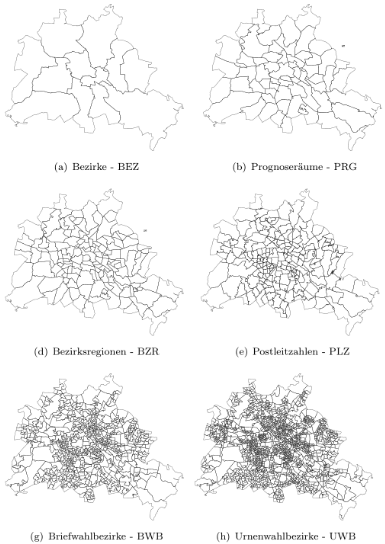
- Comparison with other map displays
  - Choropleths (standardized or normalized by area size)
  - Naive Kernel density estimates (not iterative)
  - Kernelheaping with fixed or optimum sample size
- Size of the units: The smaller the better!
- Criterion:
  - Reference: a very realistic true density
  - Bias, MSE
  - Local or average over entire region

Evaluation: A very realistic true density



Kernel density estimate of eligible voters in Berlin based on exact address information. (Source: Kerstin Erfurth (2018) Master Thesis)

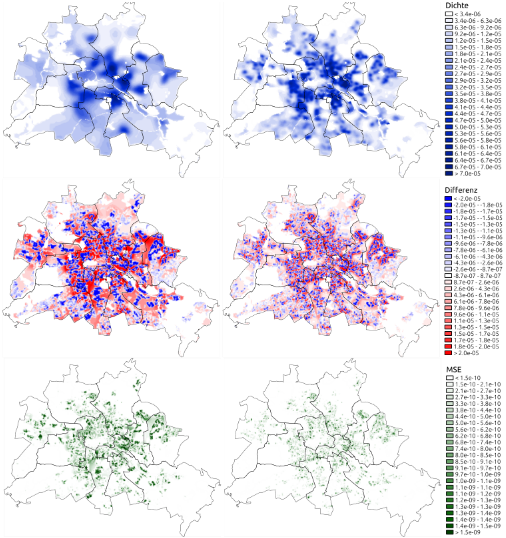
Evaluation: The size of the units



(a) Bezirke - BEZ
(b) Prognoseräume - PRG
(c) Ortsteile - ORT
(d) Bezirksregionen - BZR
(e) Postleitzahlen - PLZ
(f) Planungsräume - PLR
(g) Briefwahlbezirke - BWB
(h) Urnenwahlbezirke - UWB

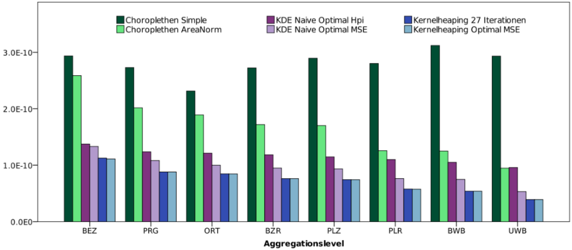
Berlin: Eight different area systems of different size.

The impact of size on the Kernelheaping Algorithm



Left: Unit= Bezirksregion (no. 4 in size scale) **Right:** Unit= Urnenwahlbezirk (no. 8 in size scale (least small))

Comparison of performance criteria



Comparison of the mean MSE for 6 different map displays and 8 different aggregation levels.

Result: Kernelheaping is the best at **all** aggregation levels! The most frequent map display is the worst and does not even improve with smaller units!

References

- Groß,M.; Rendtel, U.; Schmid,T.; Schmon,S.; Tzavidis,N. 2017: "Estimating the density of ethnic minorities and aged people in Berlin: Multivariate kernel density estimation applied to sensitive geo-referenced administrative data protected via measurement error" Journal Royal Stat. Soc. Series A , 180, 161 – 183.
- Rendtel, U.; Ruhanen, M. 2018: Die Konstruktion von Dienstleistungskarten mit Open Data am Beispiel des lokalen Bedarfs an Kinderbetreuung in Berlin. AStA Wirtschafts- und Sozialstatistisches Archiv, 12, 271–284
- Groß, M.; Rendtel, U.; Schmid, T.; Tzavidis, N. 2018: Switching between different area systems via simulated geo-coordinates: A case study for student residents s in Berlin. Discussion Paper Economics 2018/2 FB Wirtschaftswissenschaft FUB. http://edocs.fu-berlin.de/docs/receive/FUDOCS_document_000000029064
- Groß, M.; Rendtel, U.; Schmid, T.; Bömermann, H.; Erfurth, K. 2018: Simulated geo-coordinates as a tool for map-based regional analysis. Discussion Paper Economics 2018/3 FB Wirtschaftswissenschaft FUB. http://edocs.fu-berlin.de/docs/receive/FUDOCS_document_000000029065
- Erfurth, K. 2018: Gütebeurteilung und Einsatz simulierter Geokoordinaten bei der regionalen Analyse zur Bundestagswahl 2017, Master Thesis, Economic Department, Freie Universität Berlin