Erfurth, Kerstin; Groß, Marcus; Rendtel, Ulrich; Schmid, Timo

Article — Published Version

Kernel density smoothing of composite spatial data on administrative area level

AStA Wirtschafts- und Sozialstatistisches Archiv

Provided in Cooperation with:
Springer Nature

Suggested Citation: Erfurth, Kerstin; Groß, Marcus; Rendtel, Ulrich; Schmid, Timo (2021) : Kernel density smoothing of composite spatial data on administrative area level, AStA Wirtschafts- und Sozialstatistisches Archiv, ISSN 1863-8163, Springer, Berlin, Heidelberg, Vol. 16, Iss. 1, pp. 25-49, https://doi.org/10.1007/s11943-021-00298-9

This Version is available at:
https://hdl.handle.net/10419/287172

AStA Wirtsch Sozialstat Arch (2022) 16:25–49
https://doi.org/10.1007/s11943-021-00298-9

ORIGINALVERÖFFENTLICHUNG

Kernel density smoothing of composite spatial data on administrative area level

A case study of voting data in Berlin

Kerstin Erfurth · Marcus Groß · Ulrich Rendtel · Timo Schmid

Received: 1 March 2021 / Accepted: 5 December 2021 / Published online: 23 December 2021

Abstract Composite spatial data on administrative area level are often presented by maps. The aim is to detect regional differences in the concentration of subpopulations, like elderly persons, ethnic minorities, low-educated persons, voters of a political party or persons with a certain disease. Thematic collections of such maps are presented in different atlases. The standard presentation is by Choropleth maps where each administrative unit is represented by a single value. These maps can be criticized under three aspects: the implicit assumption of a uniform distribution within the area, the instability of the resulting map with respect to a change of the reference area and the discontinuities of the maps at the borderlines of the reference areas which inhibit the detection of regional clusters.

In order to address these problems we use a density approach in the construction of maps. This approach does not enforce a local uniform distribution. It does not depend on a specific choice of area reference system and there are no discontinuities in the displayed maps. A standard estimation procedure of densities are Kernel density estimates. However, these estimates need the geo-coordinates of the single units which are not at disposal as we have only access to the aggregates of some area system. To overcome this hurdle, we use a statistical simulation concept. This can be interpreted as a Simulated Expectation Maximisation (SEM) algorithm of Celeux et al (1996). We simulate observations from the current density estimates which

Kerstin Erfurth
Amt für Statistik Berlin-Brandenburg, Berlin, Germany

Marcus Groß
INWT Statistics GmbH, Berlin, Germany

Ulrich Rendtel (✉)
Freie Universität Berlin, Berlin, Germany
E-Mail: ulrich.rendtel@fu-berlin.de

Timo Schmid
Universität Bamberg, Bamberg, Germany

are consistent with the aggregation information (S-step). Then we apply the Kernel density estimator to the simulated sample which gives the next density estimate (E-Step).

This concept has been first applied for grid data with rectangular areas, see Groß et al (2017), for the display of ethnic minorities. In a second application we demonstrated the use of this approach for the so-called "change of support" (Bradley et al 2016) problem. Here Groß et al (2020) used the SEM algorithm to recalculate case numbers between non-hierarchical administrative area systems. Recently Rendtel et al (2021) applied the SEM algorithm to display spatial-temporal clusters of Corona infections in Germany.

Here we present three modifications of the basic SEM algorithm: 1) We introduce a boundary correction which removes the underestimation of kernel density estimates at the borders of the population area. 2) We recognize unsettled areas, like lakes, parks and industrial areas, in the computation of the kernel density. 3) We adapt the SEM algorithm for the computation of local percentages which are important especially in voting analysis.

We evaluate our approach against several standard maps by means of the local voting register with known addresses. In the empirical part we apply our approach for the display of voting results for the 2016 election of the Berlin parliament. We contrast our results against Choropleth maps and show new possibilities for reporting spatial voting results.

Keywords Spatial data · Administrative areas · Choropleths · Kernel density estimation · Voting atlases

Die Glättung räumlicher Datensätze auf administrativen Flächen

Eine Fallstudie mit Berliner Wahldaten

Zusammenfassung Räumliche Daten auf der Ebene administrativer Flächeneinheiten werden häufig über Karten dargestellt. Das Ziel ist es dabei regionale Unterschiede für interessierenden Bevölkerungsgruppen aufzudecken. Dies betrifft beispielsweise ältere Personen, ethnische Minderheiten, Personen mit geringer Bildung aber auch Wähler einer politischen Partei sowie Personen, die sich mit einer bestimmten Krankheit infiziert habe. Die Zusammenfassung derartiger Karten wird in Atlanten präsentiert. Eine Standarddarstellung benutzt Choroplethen, wo jede administrative Einheit durch einen einzigen Wert repräsentiert wird. Diese Karten können unter drei Aspekten kritisiert werden: Die implizite Annahme einer gleichmäßigen Verteilung innerhalb der Fläche der Einheit, die Instabilität der Darstellung beim Wechsel der administrativen Einheit sowie die Sprünge an den Grenzlinien der Einheiten, die das Aufdecken von regionalen Clustern erschweren.

Um diese Probleme zu beseitigen, verwenden wir eine Kartenkonstruktion auf der Basis von Dichten. Dieser Ansatz vermeidet eine zwangsläufige gleichmäßige Dichte innerhalb der Referenzflächen. Er ist unabhängig von der Wahl eines spezifischen Referenzsystems und vermeidet Sprungstellen. Ein Standardverfahren würde Kerndichteschätzer verwenden. Allerdings werden hierfür die Geokoordinaten der einzelnen Einheiten benötigt. Diese stehen aber nicht zur Verfügung sondern

lediglich die Aggregate der jeweiligen Flächeneinheit. Um diese Hürde zu umgehen, verwenden wir ein statistisches Simulationskonzept. Es kann als Simulierter EM (SEM) Algorithmus von Celeux et al (1996) beschrieben werden. Auf Basis der gegenwärtigen Dichteschätzung simulieren wir Beobachtungen, die mit der Aggregatsinformation konsistent sind (S-Schritt). Dann wenden wir den Kerndichteschätzer auf die simulierte Stichprobe an, die die nächste Dichteschätzung liefert (E-Schritt).

Dieses Konzept wurde erstmals für Gitterdaten auf Rechtecken zur Darstellung von ethnischen Minderheiten angewendet, Groß et al (2017). Eine weitere Anwendung fand dieser Ansatz beim sogenannten „Change of Support" Problem, (Bradley et al 2016). Hier nutzten Groß et al (2020) den SEM Algorithmus bei der Umrechnung von Fallzahlen zwischen nicht-hierarchischen Flächensystemen. Jüngst haben Rendtel et al (2021) den SEM Algorithmus für die Darstellung räumlich-zeitlicher Konzentrationen von Corona Infektionen in Deutschland verwendet.

Hier präsentieren wir drei Modifikationen des SEM Algorithmus: 1) Wir führen eine Randkorrektur ein, die die Unterschätzung der Kerndichteschätzung an den Grenzen der Population beseitigt. 2) Wir berücksichtigen unbewohnte Bereiche wie Parks, Seen und Industriegebiete bei der Berechnung der Kerndichteschätzung. 3) Wir passen den SEM Algorithmus für die Berechnung lokaler Prozentsätze an, die insbesondere für Wahlanalysen interessant sind.

Wir evaluieren unseren Ansatz gegen verschiedene Standardkarten auf Basis eines lokalen Wählerregisters mit bekannten Adressen. Im empirischen Teil wenden wir unseren Ansatz auf die Darstellung von Wahlergebnissen zur Wahl des Berliner Abgeordnetenhauses 2016 an. Wir vergleichen unsere Ergebnisse mit Choroplethenkarten und zeigen neue Möglichkeiten zur Berichterstattung räumlicher Wahlergebnisse.

Schlüsselwörter Räumliche Daten · Administrative Flächeneinheiten · Choroplethen · Kerndichteschätzer · Wahlatlas

1 Introduction

Composite spatial data are often presented by maps. The purpose of these maps is to display local clusters of subpopulations, like elderly persons, migrants, students, low educated persons, unemployed persons, persons receiving social benefits, voters of a special political party or, recently, the incidence rates of Corona infections. In most cases these maps base on count numbers for administrative area levels like federal states, counties, city districts, neighbourhoods, Zip districts or polling districts in voting analyses. Collections of thematic maps are presented in atlases,

[1] https://www.stadtentwicklung.berlin.de/planen/basisdaten_stadtentwicklung/monitoring/index.shtml (Accessed 23.9.2021)

[2] https://www.wahlen-berlin.de/wahlen/publikationen/wahlatlas/WAHLENBU_2017/atlas.html (Accessed 23.9.2021)

see, for example, the Berlin Social Indicator Atlas[1], the Berlin Voting Atlas[2], the German 2011 Zensus Atlas[3] or the EU Regional Atlas[4].

The standard maps are so-called Choropleths where the reference area is displayed by a single value, see Kraak and Ormeling (2021) (p. 170) for a recent textbook. With animated Choropleths it is possible to display additional information for the area, for example, the results of previous ballots (Tagesspiegel Wahl-Spezial (2017)). Despite its frequent use in public and scientific media Choropleth maps reveal some problems:

- The uniform representation of the reference area by a one color, which represents the area value, suggests a uniform distribution of the variable of interest within the area. This is often an unrealistic assumption.
- For different levels of aggregation, i.e. choice of administrative level, one obtains quite different maps which may lead to different conclusions.
- At the borderlines of the reference areas there are discontinuities which prevent the identification of local clusters.

These problems can be addressed by smoothing techniques, for example by Kriging, see Kriging (Oliver and Webster 2015). However, this approach uses distributional assumptions. In this paper we present a different smoothing approach which is not linked to distributional assumptions, like in the Kriging framework. The main tool is smoothing by kernel density estimation. In a first step we identify what a map should display ideally: densities or ratios of densities. As we don't observe densities we have to estimate them by kernel density estimation. However, for the kernel density estimation one needs the geo-coordinates of the units. Such information is in most cases not at hand. For example, in voting analysis one knows the geo-coordinates of the polling area at best. The exact address of the voters of a political party is to be protected for obvious reasons. In the same direction act the confidentiality rules if the data come from a survey or a register. Therefore we know only aggregate values at some area level, say, a voting district in case of ballots or an urban planning area in case of public data.

To overcome this hurdle, we use a statistical simulation concept. In an abstract view it can be interpreted as the Simulated Expectation Maximisation (SEM) algorithm of Celeux et al (1996). We simulate observations from the current density estimates which are consistent with the aggregation information (S-step). Then we apply the kernel density estimator to the simulated sample which gives the next density estimate (E-Step). The algorithm is replicated for a prefixed number of iteration after a burn-in period and the mean of the density estimates serves as the final solution.

This concept has been first applied for grid data with rectangular areas, see Groß et al (2017), for the display of ethnic minorities. In a second application we demonstrated the use of this approach for the so-called "change of support" (Bradley et al 2016) problem. Here Groß et al (2020) used the SEM algorithm

[3] https://atlas.zensus2011.de/ (Accessed 23.9.2021)

[4] https://www.destatis.de/Europa/EN/Publications/General-regional-statistics/ST_Regional_yb.html (Accessed 23.9.2021)

to recalculate case numbers between non-hierarchical administrative area systems. In the application they transferred student case numbers from Zip areas to urban planning district numbers. Recently Rendtel et al (2021) applied the SEM algorithm to display spatial and temporal clusters of Corona infections in Germany.

Here we present three adaptation of the SEM-algorithm:

- The borderline of the population area is an intrinsic problem of kernel density estimation as the standard estimates overlap the borderlines to some extent. Here we suggest to restrict the kernel functions near the borderline in an adequate fashion.
- Similarly, within a big town like Berlin there are large unsettled areas like lake, parks, industrial areas, etc. which are not settled. The simulations should respect these non-settled areas.
- Finally, ratios, like the percentage of voters for a special party, can be defined by the ratio of two densities. In this case the simulation of the samples has to be done sequentially: First the sampling of voters and then the voters of a certain party from the sample of voters.

All three adaptations are realized in the R-Package *kernelheaping* which is freely available, Groß (2021).

There are rare situations where a true realistic density is at hand to evaluate the bias and the MSE of different maps. For our analysis we got access to the geocoordinates of the Berlin voting register. From this information we could estimate a density of eligible voters, which serves as a reference value for alternative map constructions. On the basis of the register data we constructed for different aggregation levels 6 different maps (two Choropleths, two naive kernel density estimates and two versions of the SEM algorithm). We then compared the density values with the values of the reference density on a fine grid over the entire area.

Finally, we applied our approach to the results of the 2016 election of the Berlin parliament and compare it with the standard Choropleth maps. As our approach generates results which are independent from reference areas, new possibilities for spatial voting analysis arise. For example, we can compare the number of voters for a party per pixel or we can determine a highest density region for a party vote. With respect to percentages of votes we calculate the local winner at each pixel of the town.

The article is organized as follows: In Sect. 2 we introduce the density approach for the construction of maps. We then display in detail the SEM algorithm and its extensions in Sect. 3. Section 4 is devoted to the comparison of the maps in the presence of a reference density from the voters register. Finally, Sect. 5 presents the empirical analysis of the 2016 Berlin elections. Section 6 concludes.

2 A density approach for the construction of maps

2.1 Densities as the limit of area-normed Choropleths

Let the areas be indexed by $a = 1, ..., A$. For each area a the total N_a of the variable of interest is known. The total number of cases in the population N is given by $\sum_{a=1}^{A} N_a$ Furthermore, let Δ_a be the size of area a.

A naive version of a Choropleth maps uses the value N_a as area value. However, this version has the severe disadvantage that large areas are regularly over-represented, see Kraak and Ormeling (2021). A better solution is the use of N_a/Δ_a as area value, which is the number of observation per area unit. We call it an area-normed Choropleth. Here the integral over the Choropleth map results in the total number of cases N over the entire region. If we decrease the size of the reference areas the limit $1/N \times N_a/\Delta_a$ will become the density f of the variable of interest at the spot $x = (x_1, x_2)'$ where the area a is concentrated. Thus the density $f(x_1, x_2)$ is the natural generalisation of the area-normed Choropleth map. Note, that maps which display levels of the density $f(x_1, x_2)$ are independent from aggregation levels. There is no build-in discontinuity and if the density is constant over a certain region, then the distribution of the variable of interest is uniform within that region. Thus the use of densities solves the above mentioned problems of Choropleth maps.

Of course we do not know the density f and therefore we have to estimate it. A well-known estimator is the kernel density estimator $\widehat{f}$ (Härdle 1991):

$$\widehat{f}(x) = \frac{1}{N|H|} \sum_{k \in U} K\left(H^{-1}(x_k - x)\right), \tag{1}$$

where K is the kernel function, H is a symmetric positive definite bandwidth matrix and $|\cdot|$ denotes the determinant. The selection of the bandwidth is important for the performance of the kernel density estimator (1). However, as the main focus here is not on the selection of bandwidth we use the plug-in approach proposed by Wand and Jones (1994) and set $H = \text{diag}(h_1, h_2)$ with suitably chosen smoothing parameters h_1 and h_2. A common choice for K, used in this paper, is the Gaussian Kernel function $K(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2} x'x)$.

To compute the kernel density estimate it is necessary to know the geo-coordinates of units. This unrealistic assumption will be relaxed in the next section.

2.2 The estimation of local proportions

Often Choropleth area counts are normed by a second variable, for example, the number of voters for a party among all voters. In this case the Choropleth converges to a ratio of two densities, the density of voters of a party and the density of voters.

To see this, let f_V be the density of voters. Correspondingly let f_P be the density of voters of party P. Furthermore, let N_V be the total number of voters and let N_P the total number of voters for party P. The expected number of voters at an rectangle of size $\Delta_{x_1} \times \Delta_{x_2}$ at coordinate $x = (x_1, x_2)'$ is approximately given by $N_V \cdot f_V(x_1, x_2) \cdot (\Delta_{x_1} \times \Delta_{x_2})$. Similarly, the expected number of voters for party P

at coordinate $x = (x_1, x_2)'$ is obtained by $N_P \cdot f_P(x_1, x_2) \cdot (\Delta_{x_1} \times \Delta_{x_2})$. Hence, the ratio

$$r(x_1, x_2) = \frac{N_P}{N_V} f_P(x_1, x_2) / f_V(x_1, x_2)$$

has the interpretation of a local percentage of voters for party P, which corrects the population average $\frac{N_P}{N_V}$ to the local level.

A standard nonparametric estimator of local ratios $r(x)$ is the Nadaraya-Watson estimator $\widehat{r}_{NW}$, see (Härdle 1991). The estimator can be shown to be the ratio of two kernel density estimates with an equal smoothing factor. To see the equivalence in our example let U_V be the universe of voters and N_V total number of voters. Similarly we obtain for party P voters U_P and N_P. Let P_k denote the a dummy variable, which indicates whether voter k is a voter of party P ($P_k = 1$) or not ($P_k = 0$). The Nadaraya-Watson estimator $\widehat{r}_{NW}$ is then given by:

$$\begin{aligned} \widehat{r}_{NW}(x) &= \frac{\frac{1}{N_V}\sum_{k \in U_V} \frac{1}{|H|} K\left(H^{-1}(x - X_k)\right) P_k}{\frac{1}{N_V}\sum_{k \in U_V} \frac{1}{|H|} K\left(H^{-1}(x - X_k)\right)} \\ &= \frac{N_P \widehat{f}_P(x)}{N_V \widehat{f}_V(x)} \end{aligned} \tag{2}$$

where the last line is the scaled ratio of the kernel density estimates of the density of the party and the density of voters.

As the number of voters for a party is smaller than the number of voters it is reasonable to select the smoothing factor of the party distribution which is generally somewhat larger than the corresponding value of the voters distribution.

3 The SEM algorithm for the estimation of densities

3.1 The baseline SEM algorithm

Now we describe the SEM algorithm for the estimation of the kernel density estimate $\widehat{f}$.

To keep things numerically tractable we generate x-coordinates only on a fine grid of geo-coordinates and we evaluate the resulting density estimate only on these grid-points. Let x_g $(g = 1, ..., G)$ be the geo-coordinate of the G grid points. Then the set $\mathcal{G} = \{x_g | g = 1, ..., G\}$ can be separated into A subsets $\mathcal{G}_a$, where all members belong to area a. The double indexed $x_{g,a}$ displays the geo-coordinate of grid point g belonging to area a. We assume that the area centroids y_a are known for all units k in the universe U_a of area a.

The basic SEM algorithm may be formulated as follows:

Step 1 Compute an initial kernel density estimate $\widehat{f}^{(0)}$.

- Use $x_k^{(0)} = y_a$ for all $k \in U_a$.
- Set the smoothing parameters $h_1^{(0)}$ and $h_2^{(0)}$ to sufficiently large values such that no spikes occur in the density estimate.
- Calculate $\widehat{f}^{(0)}(x)$ for all $x = x_{g,a}$ for all $g = 1, ..., G$ and all $a = 1, ..., A$.

Step 2 Draw a stratified sample $s^{(n)}$ from $\{x_{g,a} | g = 1, ..., G;\ a = 1, ..., A\}$.

- The strata sizes are N_a $(a = 1, ..., A)$.
- The sampling is with replacement. The sampling weights are proportional to $\widehat{f}^{(n-1)}(x_{g,a})$ as size variable.
- The sampling size in the stratum of area a is N_a.

Step 3 Recalculate $\widehat{f}^{(n)}$ from the sample $s^{(n)}$.

- Determine the smoothing parameters $h_1^{(n)}$ and $h_2^{(n)}$ by the plug-in estimator of Wand and Jones (1994). Note that other selectors for the bandwidth matrix H can be also applied.
- Calculate $\widehat{f}^{(n)}(x)$ for all $x = x_{g,a}$ for all $g = 1, ..., G$ and all $a = 1, ..., A$.

Step 4 Repeat Steps 2 and 3 B times for a burn-in phase and R times for replication.
Step 5 The final density estimate $\widehat{f}(x)$ is:

$$\widehat{f}(x) = \frac{1}{R} \sum_{r=1}^{R} \widehat{f}^{(B+r)}(x).$$

This algorithm can be realized with the *R*-Package *kernelheaping* (Groß 2021).

3.2 The boundary correction for unsettled areas

25% of the area of Berlin are lakes, forests, parks, industrial areas which are not settled. So the kernel density estimate should not cover these areas[5]. A straightforward approach to this problem is to restrict the kernel function to the settled area and to rescale it to a probability function by a suitable constant, see Jones (1993). Note, that the rescaling factor varies for each point on the grid.

The rescaling approach basically controls which part of the kernel function lies within the settlement area $\mathcal{S}$. For this purpose one has to compute for every coordinate x the weight:

$$w_x = \int_{\mathcal{S}} \frac{1}{|H|} K\left(H^{-1}(x - y)\right) \mathrm{d}y. \tag{3}$$

Note, that the weight w_x depends on the smoothing parameters h_1 and h_2.

[5] This should also hold for Choropleth maps. Although it is quite simple to exempt unsettled areas from these maps it is not practised in voting analyses, see Amt für Statistik Berlin-Brandenburg (2016).

The rescaled kernel density estimate $\widehat{f}_{rs}(x)$ at geo-coordinate x is then given by:

$$\widehat{f}_{rs}(x) = \frac{1}{N|H|} \sum_{k \in \mathcal{S}} \frac{1}{w_x} K\left(H^{-1}(x - x_k)\right). \tag{4}$$

In the discrete setting of the grid $\mathcal{G}$ the grid points which lay inside $\mathcal{S}$ are denoted by $\mathcal{G}_S$. Furthermore, let $\Delta_{\mathcal{G}}$ be the area between four neighboring grid points. Then, the weight w_x at coordinate x can be approximated by

$$w_x \approx \sum_{z \in \mathcal{G}_S} \frac{1}{|H|} K\left(H^{-1}(x - z)\right) \Delta_{\mathcal{G}}. \tag{5}$$

In the case of a Gaussian Kernel we obtain:

$$w_x = \frac{\Delta_{\mathcal{G}}}{\sqrt{2\pi}} \frac{1}{h_1 h_2} \sum_{(z_1, z_2) \in \mathcal{G}_S} \exp\left\{-0.5\left(\frac{(x_1 - z_1)^2}{h_1} + \frac{(x_2 - z_2)^2}{h_2}\right)\right\}, \tag{6}$$

and w_x is computed for every $x \in \mathcal{G}_S$. As the number of grid points increases in a quadratic fashion with the grid length, the computation of the w_x may turn out to be computer intensive as the weights w_x have to be recalculated in every iteration step of the SEM algorithm because they depend on the bandwidth matrix H. The modified SEM algorithm which computes the rescaled kernel density estimate $\widehat{f}_{rs}$ can be found in the Appendix A. It is also implemented in the R-package *kernelheaping* (Groß 2021).

3.3 The estimation of local proportions

As shown above, the Nadaraya/Watson estimator can be computed as the ratio of the two kernel density estimates of the party voters and the voters. For the simulation of the corresponding densities we have to consider that the party voters are a subset of the voter. Hence the selection of the sample of party voters—and their coordinates—has to be taken from the sample of voters.

The corresponding algorithm can be found in Appendix B and it is implemented in the R-package *kernelheaping* (Groß 2021).

4 Evaluation study

In this section we present results of a validation study for assessing the performance of the proposed SEM algorithm and alternative map presentations. The aim is to investigate the ability of the proposed SEM algorithm to deal with aggregated information and hence provide more accurate estimates than alternative standard map presentations. The evaluation of the proposed algorithm is based on a list of all addresses in Berlin in December 2016 which is 3 month after the election of September 2016 which we analyze in Sect. 5. In Fig. 1 every dot represents a valid address in

Fig. 1 Distribution of addresses in Berlin

Berlin. The white areas represent unsettled areas. Now, a number of eligible voters lives at every address. This number can change considerably between addresses. For privacy reasons the true number of eligible voters has been slightly changed by adding a small random component by the data provider. With this information we estimated a kernel density function—which respects the boundaries of unsettled areas and of Berlin—on a 100 m × 100 m grid. Figure 2 displays this *reference/true density* in the evaluation study. The colors are scaled to the number of eligible voters per 100 m × 100 m. This area corresponds to a pixel of the screen representation.

In order to assess the performance of the proposed algorithm we aggregate the eligible voters at their addresses according to 8 different (aggregation) area levels. The highest level BEZ (Bezirke), is defined by 12 Berlin city districts. The next lower levels PRG (Prognoseräume) are 60 major prediction areas followed by 96 ORT (Ortsteile) city parts. The next stages are given by 136 BZR (Bezirksregion) district areas, 192 PLZ (Postleitzahl) Zip code areas and 447 PLR (Planungsräume) planning areas. The most fine area systems are closely related to the voting regulations. The voters have the possibility to vote by letter or to go to a place where they can put their vote into a bin, the urn. In Berlin there are 600 BWB (Briefwahlbezirke) postal voting districs and 1779 UWB (Urnenwahlbezirke) ballot voting districts. Figure 3 displays the granularity of these area systems. Note, that these area systems not not hierarchically ordered.

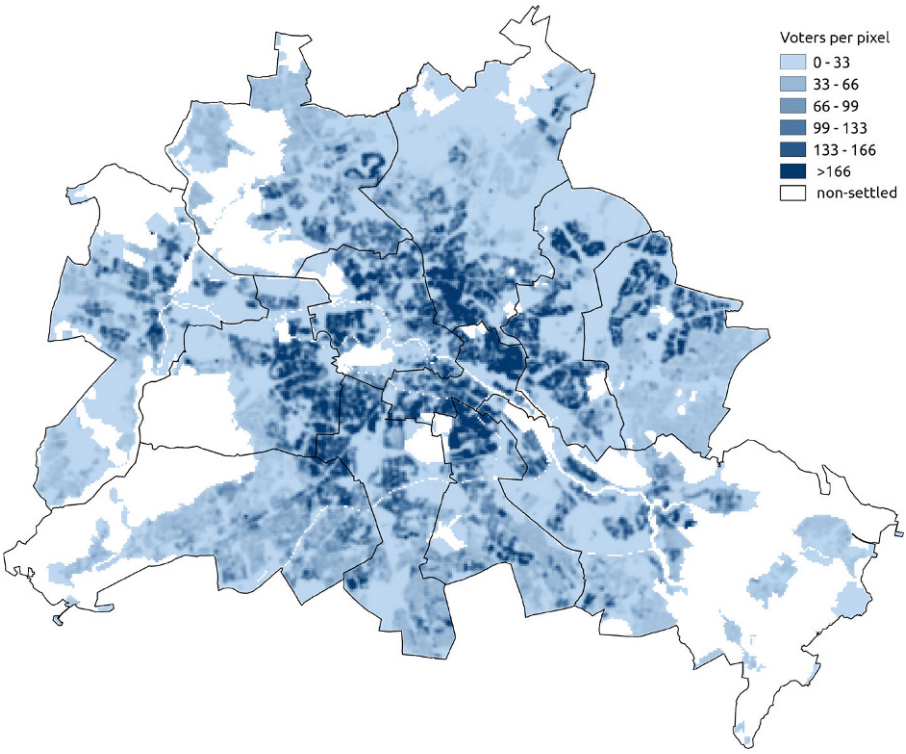

Fig. 2 Distribution of eligible voters in Berlin (December 2016)

We compare 6 different map representations with the *reference/true density* in Fig. 2. The first two are linked to Choropleth representations. In the first version the area value of the Choropleth is given by the number of voters in the area[6] (denoted by Choropleth Simple). The second version of Choropleth maps divides the area count numbers by the size of particular area[7] (denoted by Choropleth AreaNorm). Both versions are normalized by a constant to a density such that the results can directly be compared to the *reference/true density* in Fig. 2. However, the interpretation of the scaled Choropleths remains unchanged.

Furthermore, we use two non-iterative kernel density estimators (with different smoothing parameters) in the simulation. In both versions the centroid of the area is used as the geo-coordinate for the estimation. In the first version we use the smoothing parameter which is derived from the *reference/true density*density (denoted by KDE Naive Optimal TRUE). As the *true density* is in general unknown we use in the second version the optimum smoothing parameter for the current sample (denoted by KDE Naive Optimal Sample).

[6] Note, that we did not group these numbers into intervals. Thus results for these Choropleth maps are somewhat more informative compared to its grouped version.

[7] Note, that we use here only the settled area. This makes the map more realistic than the standard use which ignores unsettled areas.

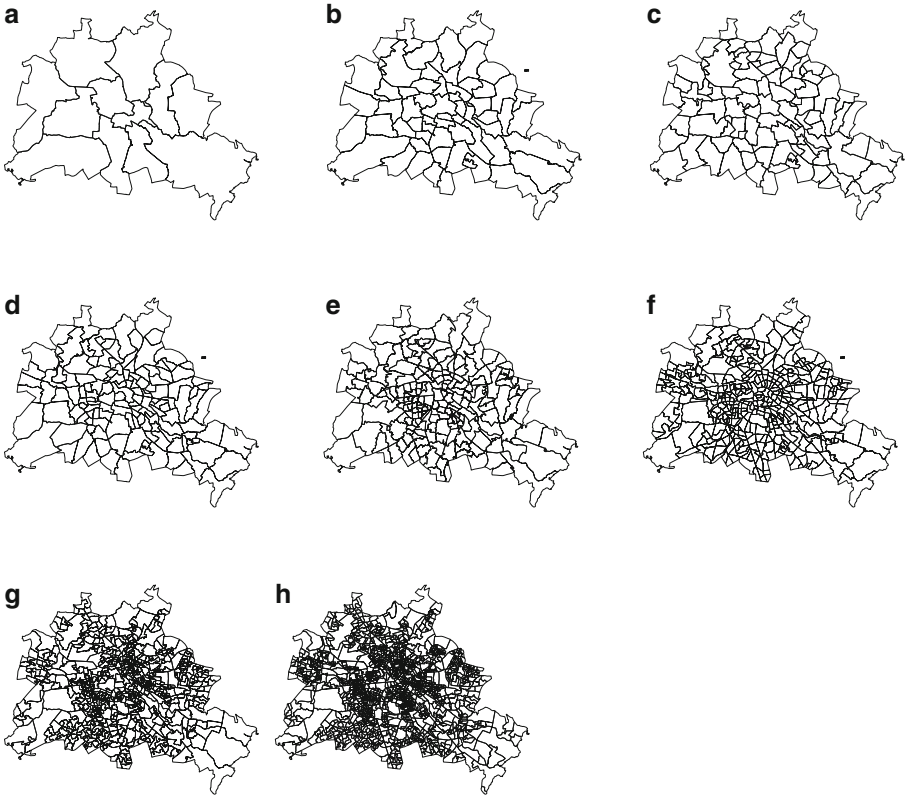

Fig. 3 Comparison of the granularity of 8 area systems in Berlin. **a** City districs—BEZ. **b** Prediction areas—PRG. **c** City parts—ORT. **d** District areas—BZR. **e** Zip areas—PLZ. **f** Planning areas—PLR. **g** Postal voting districs—BWB. **h** Ballot voting districs—UWB

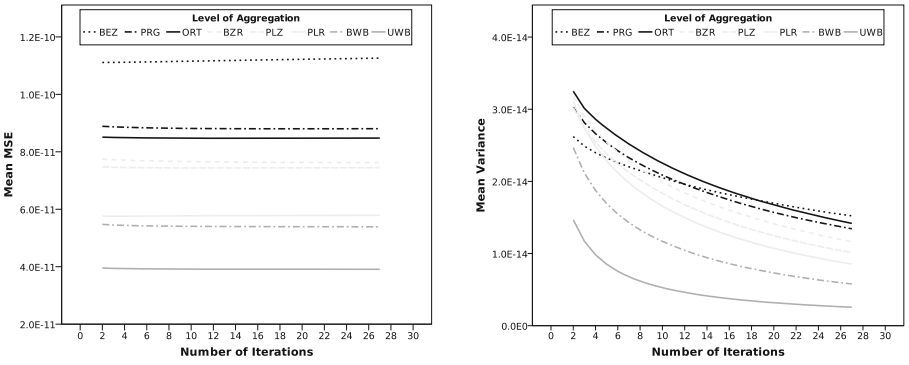

Fig. 4 Comparison of MSE and variance of the *kernelheaping* procedure for different aggregation levels and iterations

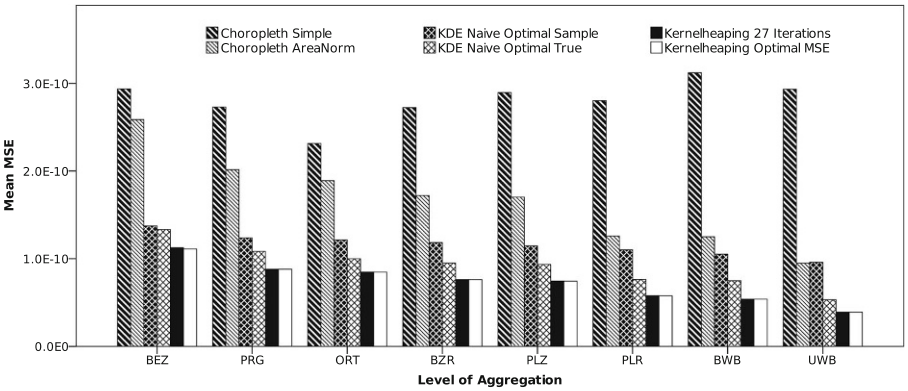

Fig. 5 Comparison of mean MSE for 6 map constructions and 8 aggregation levels

The remaining two map presentation are based on the proposed SEM algorithm.The most sensitive parameter for the SEM algorithm is the number of iterations after the burn-in phase. As shown in Fig. 4 the mean MSE is quite sensitive with respect to the selected area level: the lower the area level the lower is the mean MSE. However, the number of iterations after the burn-in phase has a low impact on the mean MSE. This is due to the small size of the variance component[8], which amounts only a factor 10^{-3} of the MSE, see Fig. 4 in the right panel. Thus, the main contribution of the MSE is the bias component. For the comparison of the MSE with the other maps we use two versions of the SEM algorithm. In the first version we use a constant number of replications, which is set to $R = 27$ (denoted by Kernelheaping 27 Iterations). In the second version the number of replications is optimized such that the MSE is minimized (denoted by Kernelheaping Optimal MSE).

Figure 5 compares the mean MSE of the 6 map constructions over the 8 different levels of aggregation. With the exception of the simple Choropleth map all maps improve if a lower level of aggregation is chosen. The area-normed choropleth map performs reasonably well at a very low aggregation level. Remember, however, that the MSE for all Choropleth versions is too optimistic as we ignored the grouping of area values and the standard ignorance of unsettled areas in applications. The naive kernel density estimate with a fixed smoothing parameter which is selected by knowledge of the true density performs well for low aggregation. The SEM algorithm performs best at all levels and the MSE is quite robust against the number of replications.

Having assessed the mean MSE of the different map representations we investigate the visual impression of the corresponding maps. Therefore, the Figs. 6 and 7 display the resulting maps for the simple Choropleth map and the kernelheaping map for a high (138 district areas BZR) and a low (1779 ballot voting districts UWB) level of aggregation. Additionally, we display the over- (Color Red) and

[8] The low absolute size of this component is also due to the fact that we displayed here variances instead of standard deviations.

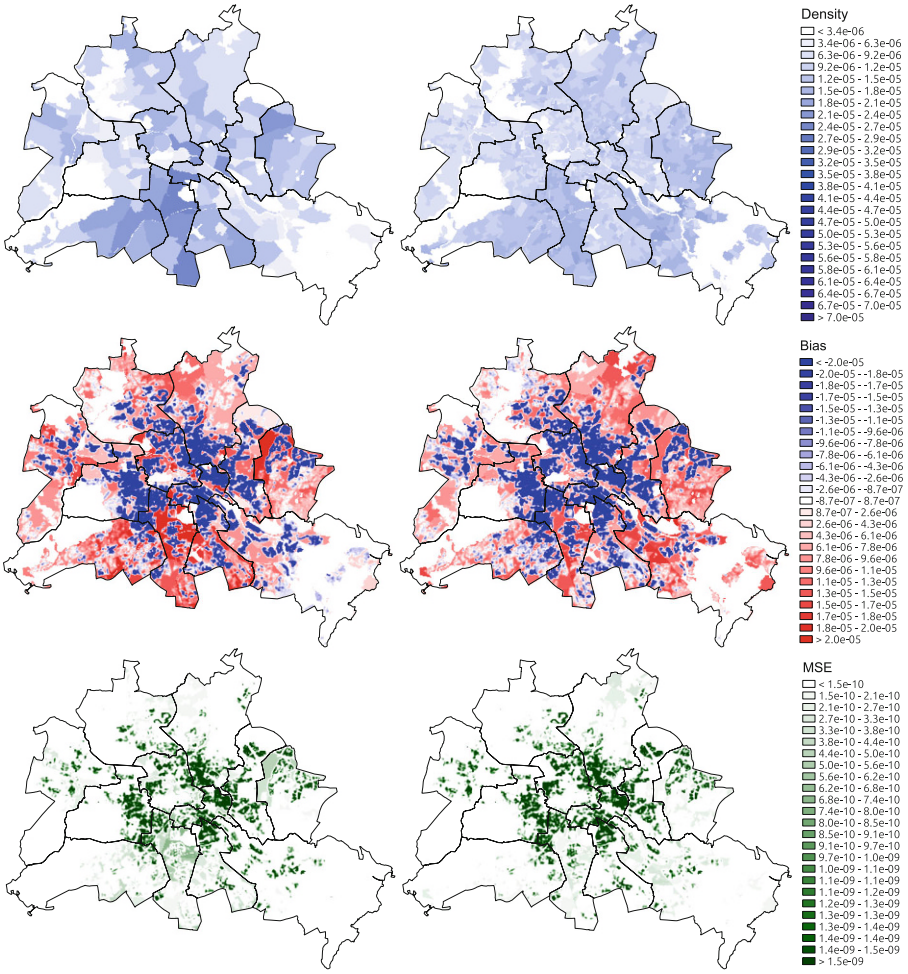

Fig. 6 Density estimates, bias and MSE (top down) of the simple Choropleth map for two different levels of aggregation: district area (BZR, *left panel*) and ballot voting district (UWB, *right panel*). *Blue* under-estimation of *reference/true density*. *Red* over-estimation of *reference/true density*

under-estimation (Color Blue) of the true density. In the third row the local MSE values are displayed. In order to ease comparisons, the scale of the figures for the Choropleth map and the kernelheaping map are identical.

The Choropleth maps in Fig. 6 do by no means reflect the structure of the voter population in Berlin. Even at the smallest possible level of aggregation (UWB level—right panel) one has the impression that the voters are equally spread over the city, with the exception of the unsettled areas. On the contrary, the kernelheaping maps in Fig. 7 reflect well the dense voter belt which surrounds the center of the town. This is seen even at a fairly high aggregation level (BZR level—left panel). The impression becomes even more informative when we go to the lower aggregation level (UWB level—right panel).

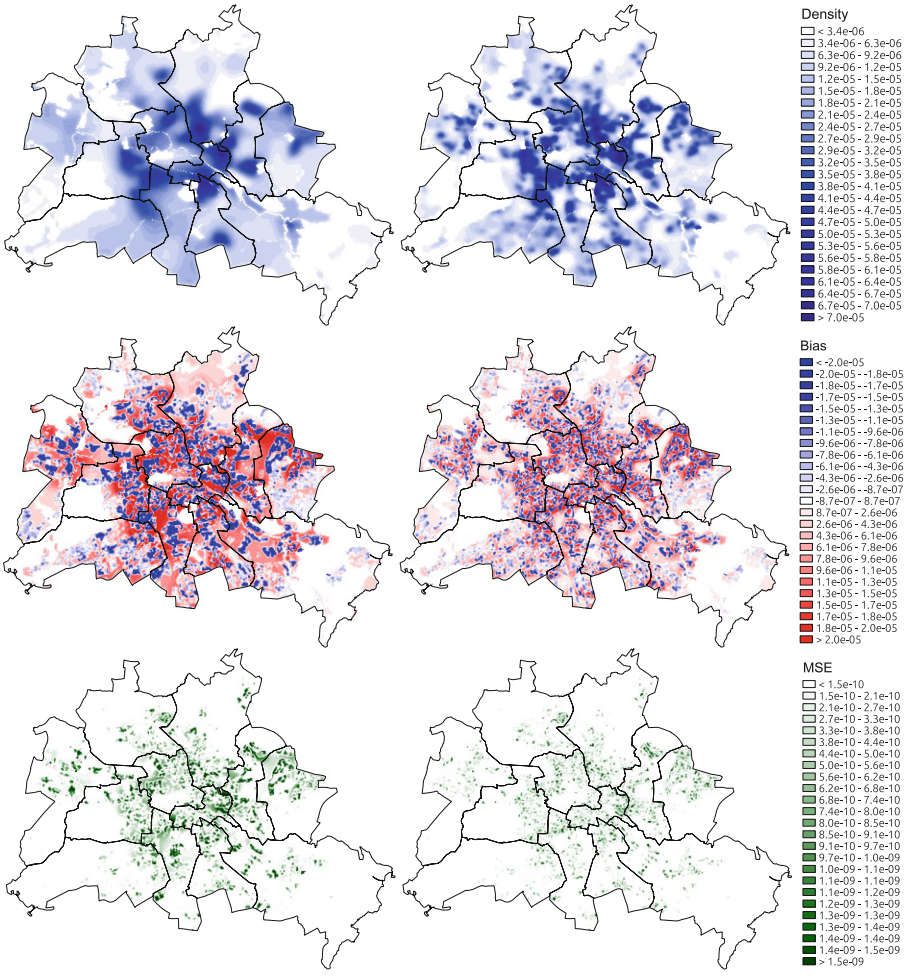

Fig. 7 Density estimates, bias and MSE (top down) of the Kernelheaping map for two different levels of aggregation: district area (BZR, *left panel*) and ballot voting district (UWB, *right panel*). *Blue* under-estimation of *reference/true density*. *Red* over-estimation of *reference/true density*

Note the high resolution of the *reference/true density* which displays even larger streets. Of course, such specific features will be ignored by the Choropleth maps and even by the kernelheaping maps and therefore for these areas the resulting map over-estimates the voter density. One might object that such a high resolution is not the object of a substantive voting analysis.

Finally, if we compare the second with the third row of Figs. 6 and 7 we see that the regional distribution of the MSE is determined to a large extent by the regional bias.

5 Application to Berlin voting data

5.1 Number of voters per pixel

We display the application of the technique of simulated geo-coordinates for the results of the general election of the Berlin regional parliament in 2016. The data are freely available under the link https://www.wahlen-berlin.de/Wahlen/BE2016/afspraes/download/download.html. Special emphasis is given to the results for the AfD, a new right wing party in the spectrum of German political parties. At this election the overall percentage for the AfD was 14.1%.

In a first step we look for the regional distribution of AfD-voters. The densities for the distribution of voters are normalized to a volume of 1 under their surface. In order to make them comparable they should be multiplied by the absolute number N_P of voters for party P. If we multiply the densities with the area of the pixels of the maps, which is $140 \times 140\,\mathrm{m}^2$ in our case, we end-up with a scale which can be interpreted as the number of voters of party P per pixel.

Figure 8 compares for the AfD the results of the re-scaled density maps with the Choropleth representation. Both maps exclude unsettled areas of Berlin. There are striking differences in the regional distribution suggested by the maps. Even with the exclusion of the unsettled areas of Berlin the Choropleth representation suggests a strong AfD frequency in the south east of Berlin which is not confirmed by the density representation. According to the density map there is a sizeable concentration of AfD voters in the very east of Berlin. The map also indicates reasonable concentrations of AfD voters in the former West-Berlin part of the town. This is not recognized from the Choropleth map.

One of the most powerful features of the kernel density approach is the characterization of clusters by high density areas. Figure 9 displays the high density area for AfD voters. The displayed area covers 20% of all AfD voters based on the proposed SEM algorithm. Within these clusters the density is larger than 12 voters per pixel. The area is split into single regional clusters. Most of the clusters represent city

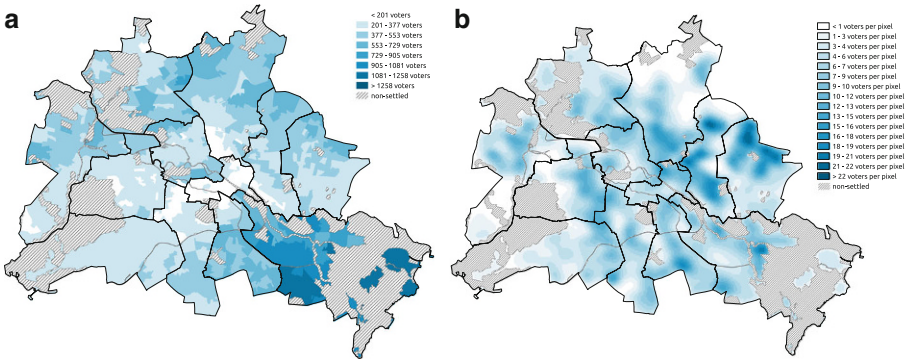

Fig. 8 Number of voters for party AfD in regional elections 2016 in Berlin. Absolute number displayed by simple Choropleth on the level of (postal) voting districts (**a**) and the number of voters per pixel ($=$ $140 \times 140\,\mathrm{m}^2$) via kernelheaping map (**b**)

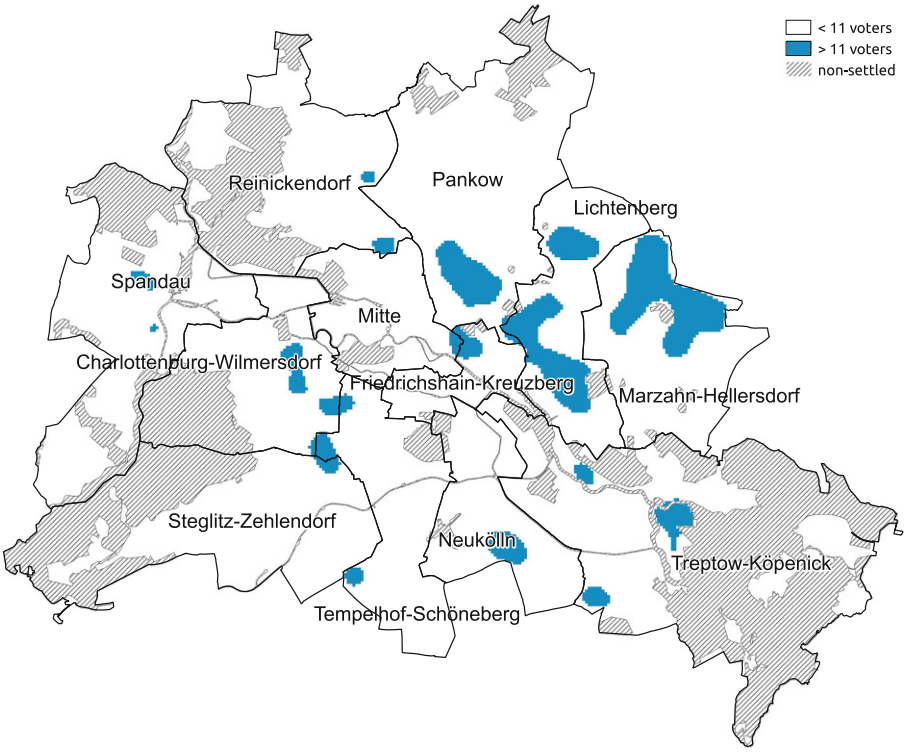

Fig. 9 High density area covering 20 percent of AfD voters

quarters with tower building flats from the 70-s to the 90-s of the last century. This does not only hold for the former East-German settlements in the district Marzahn-Hellersdorf but also for the former West-Berlin settlements Gropius-Stadt in the south of the district Neukölln and the Märkisches Viertel in the east of the district Reinickendorf. Such an identification of regional clusters is a good starting point for an analysis of voting behaviour. Note, that these clusters cannot be identified from the Choropleth map of Fig. 8.

A different attractive feature is the comparability of the re-scaled densities for different parties. So one can display for each point the party which achieves the highest number of voters per pixel. Figure 10 displays the best areas per pixel for the Christian-Democrats (CDU in dark blue), the Social-Democrats (SPD in red), the GREEN party (Grüne in green), the Left-Wing Party (Linke in purple) and the already mentioned AfD (AFD in light blue).

5.2 The analysis of local percentages

If we switch to the estimation of local percentages we first have to estimate the distribution of the voters. Figure 11 displays a density estimate of the distribution of

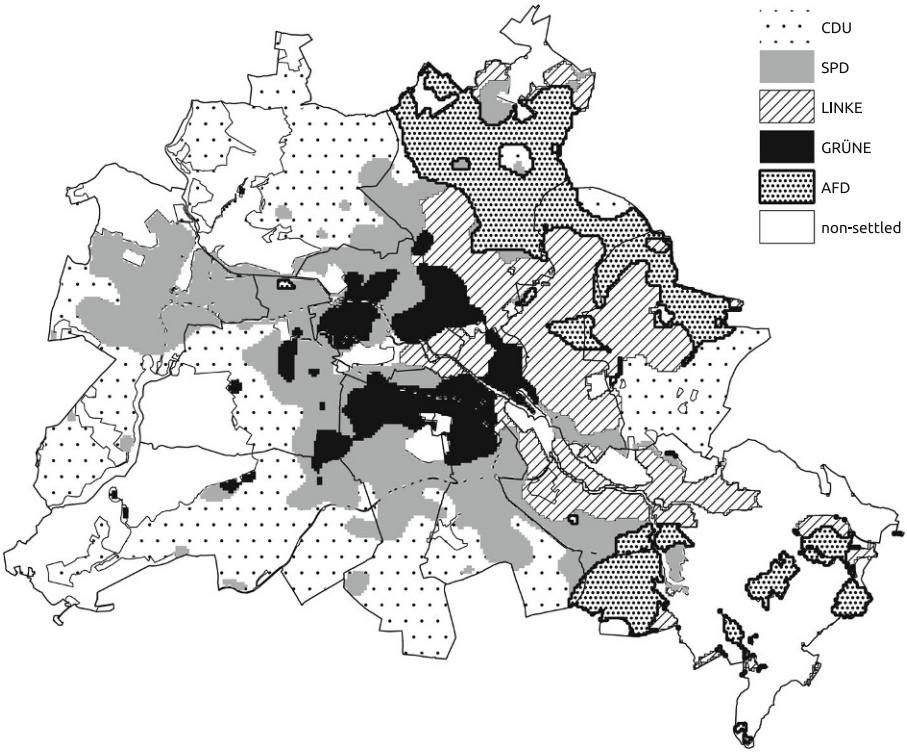

Fig. 10 The winner with respect to the highest number of voters per pixel ($= 140 \times 140\,\mathrm{m}^2$)

voters[9] per pixel. This density varies considerably within Berlin which is the reason why the Choropleth maps of absolute figures are misleading in this case.

Figure 12 compares the local proportions of AfD voters via density estimation with the proposed SEM algorithm with the percenatges in (postal) voting districts. There is a high coincidence of results between the two maps, displaying high percentage numbers in the south-east and the north-east of Berlin. However, the map of the percentages in single voting districts is more erratic and exhibits adjacent voting districts with low and high percentages.

With the local percentage it is possible to create two versions of high percentage areas. The first version asks for the area where a prefixed limit is exceeded. Such an area is shown in Fig. 13 for a limit of 10 percent for the AfD. It displays for broad regions a substantial support of the AfD.

The second possibility to display high percentage areas is to keep the percentage of the covered area fixed, say 20 percent of the Berlin area, and to ask for the limiting percentage which defines the borderline of this area. Such a display is convenient for comparisons between different parties. Figure 14 compares the high percentage areas for the six parties which became elected into the parliament. For each party the

[9] In the evaluation section we analyzed the distribution of **eligible** voters while we here analyze the distribution of persons who really voted.

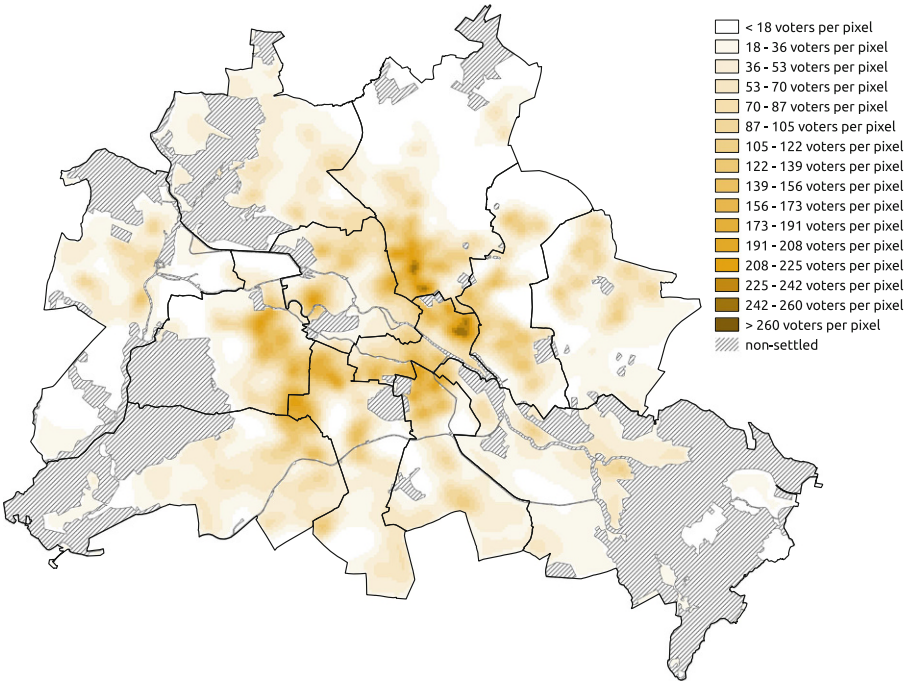

Fig. 11 The number of voters per pixel ($= 140 \times 140\,\mathrm{m}^2$)

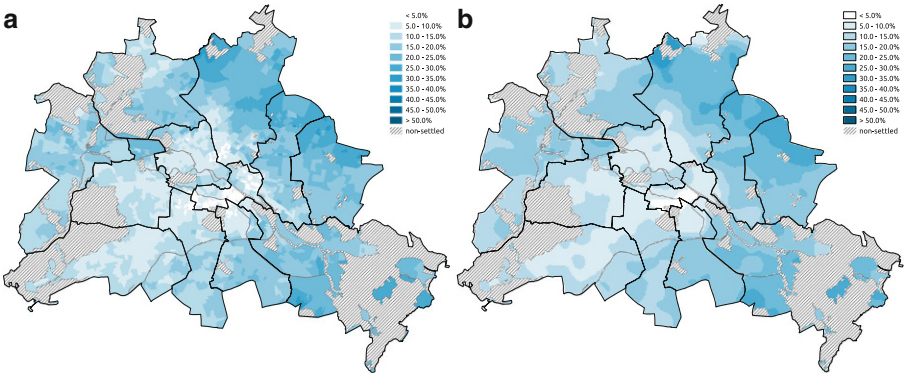

Fig. 12 Percentage of AfD-Voters: Proportions in voting districts (**a**) and local proportions via densities (**b**)

covered part of the settled area of Berlin is 20 percent. However, the party specific areas cover quite different parts of Berlin. For example, the right wing AfD and left wing LINKE are almost entirely concentrated on the former East-Berlin. Also the limit values, which define the borderline of the areas, vary substantially. Table 1 compares these limit values with the average percentages of the party at the Berlin level. By definition the limit value is higher than the average over Berlin. However, the difference between these baseline figures are small for the SPD and the GRÜNE

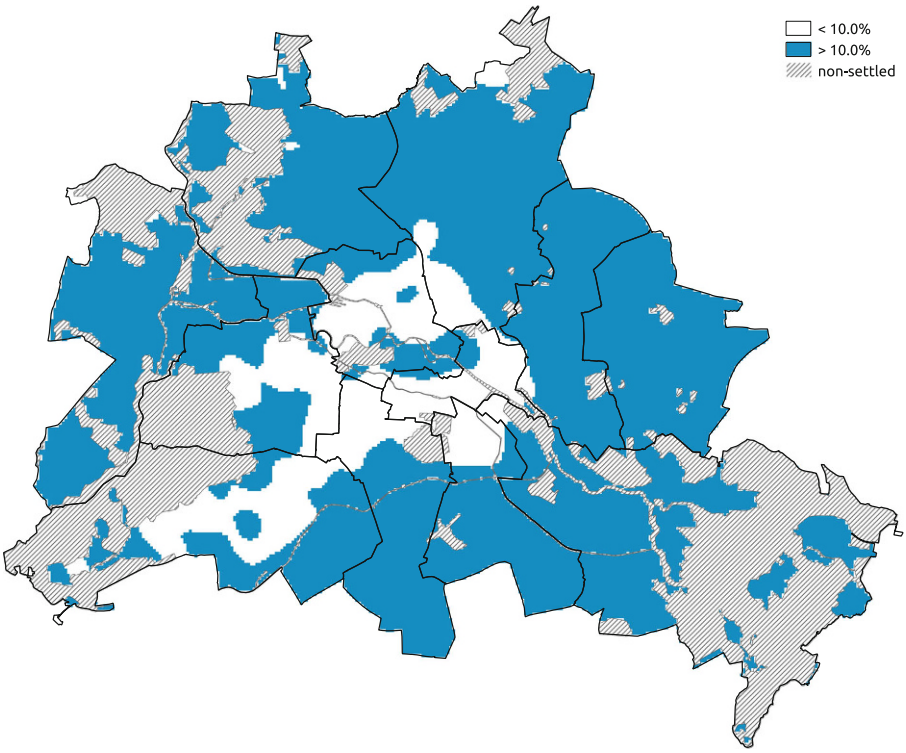

Fig. 13 High percentage areas: Percentage for AfD is larger than 10%

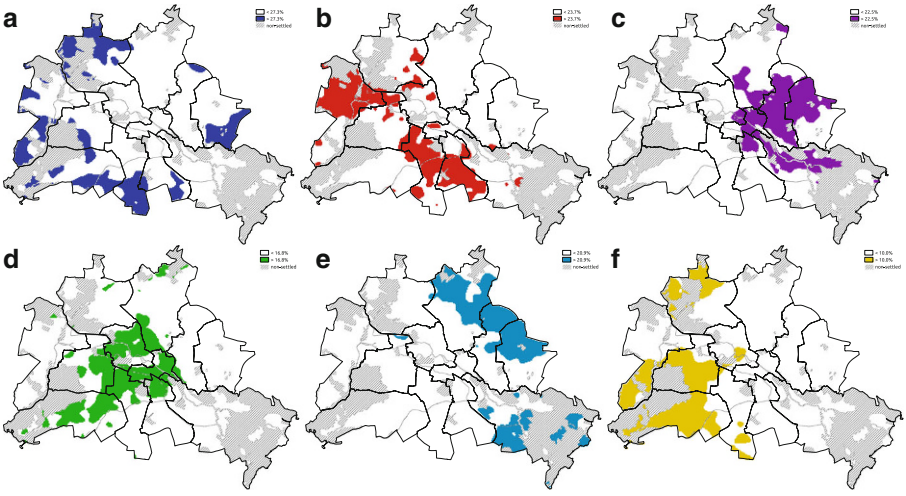

Fig. 14 High percentage areas for 6 parties: **a** CDU (*dark blue*), **b** SPD (*red*); **c** Linke (*purple*), **d** Grüne (*green*), **e** AfD (*light blue*), **f** FDP (*yellow*). Covered area is 20% of the settled population

Table 1 Comparison of the limit values of high percentage areas and the average percentage over the Berlin area for different parties

Party	Limit value of area	Average value Berlin
CDU	27.3	17.3
SPD	23.7	21.6
LINKE	22.5	15.6
GRÜNE	16.8	15.2
AfD	20.9	14.2
FDP	10.0	6.7

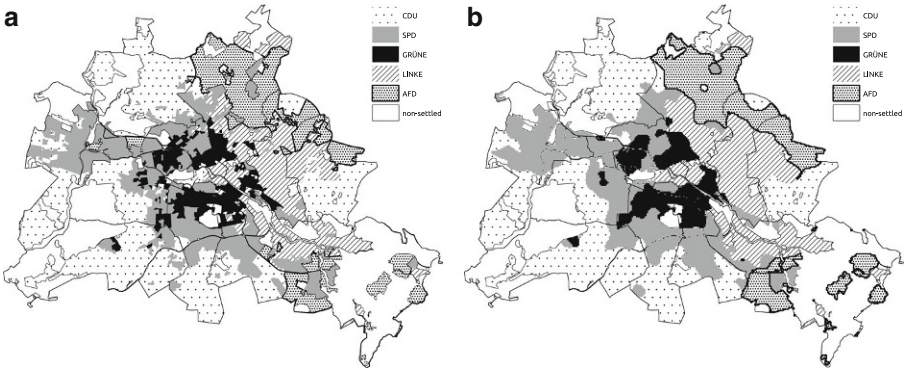

Fig. 15 The winner of the voting districts (**a**) compared to the local percentage winner (**b**)

party and they are much bigger in the case of the other parties. This indicates that the results for the SPD and the GRÜNE party are more homogeneously distributed than for other parties.

Finally, local percentage maps offer the possibility to display at each point of the city the party with the highest percentage. Because of the smooth shape of the local percentages their maximum is also smooth. Figure 15 compares a map of the local majority derived from the densities (right) with a Choropleth which displays for each voting district the color of the party with the maximum percentage in the district (left). Despite the different construction the two maps give a similar impression where the respective parties have a local majority.

6 Concluding remarks

It is the aim of a spatial analysis to link information on local concentrations with regional information from other sources. In the previous examples we used information about the former division of Berlin into East- and West-Berlin. We also used information about the settlement structure of Berlin. Such additional information can be displayed by background maps which can be combined with the density maps. Such an enrichment of maps with information is the general aim of GIS-software, see the textbook of Mitchell (2005) on Spatial Measurement and Statistics.

The approach presented here can be applied to any composite spatial data on administrative levels. In our example we used official voting records at different aggregation levels. Often the local aggregates can be accessed via an open data portal; for example, the open data portal of Berlin may be reached via the link https://daten.berlin.de/. Rendtel and Ruhanen (2018) used spatial demographic data from the open data platform to construct a map of child density and compared the density of children with the allocation of kindergardens and pediatrists in Berlin to assess the local fit of needs and offer.

If the data come from a survey we may either use the estimated totals for the spatial areas at some level or we may use the survey data directly. In this case we will have to use the survey weights. The procedure *kde* for the kernel density estimation from the R-package *ks* which is used for the *kernelheaping* package can deal with survey weights. However, there is no special input parameter for a vector of survey weights in *kernelheaping*. This has to managed by the user of the *kernelheaping* package.

A display of the precision of the densities and proportions is rarely found in standard maps. If the aggregates come from registers and official sources there is no need to do this because there is no statistical variation, at least theoretically. However, the SEM-algorithm has a stochastic component: the repeated sampling from the estimated densities. In this case the variance can be easily determined from the variance of the replicates, see Groß et al (2020). However, a variance component which is due to sampling is not jet covered by the *kernelheaping* package.

Appendix

The modified SEM Algorithm with boundary correction

Step 1 Compute the initial kernel density estimation $\widehat{f}_{rs}^{(0)}$:

- Use $x_k^{(0)} = y_a$ for all $k \in U_a$: All units are supposed to lay in the settled area $\mathcal{S} \subset U$. Also the area centroids are supposed to lay in settled areas. The computation of the centroids may be affected by the exemption of the unsettled areas from the original areas.
- Set the smoothing parameters $h_1^{(0)}$ and $h_2^{(0)}$ to sufficiently large values such that no spikes occur in the density estimate.
- Compute weights $w_x^{(0)}$ for every $x \in \mathcal{G}_S$, the set of gridpoints in the settled area.
- Calculate $\widehat{f}_{rs}^{(0)}(x)$ for all $x \in \mathcal{G}_S$.

Step 2 Draw a stratified sample $s^{(n)}$ from $\mathcal{G}_S$.

- The strata sizes are N_a $(a = 1, ..., A)$.
- The sampling is with replacement. The sampling weights is proportional to $\widehat{f}_{rs}^{(n-1)}$ as size variable.
- The sampling size in the strata of area a is N_a.

Step 3 Recalculate $\widehat{f}_{rs}^{(n)}$ from sample $s^{(n)}$.

- Determine the smoothing parameters $h_1^{(n)}$ and $h_2^{(n)}$ by the plug-in estimator of Wand and Jones (1994). Note, that other selectors for the bandwidth matrix H can be also applied.
- Determine adapted weights $w_x^{(n)}$ for every $x \in \mathcal{G}_S$.
- Calculate $\widehat{f}_{rs}^{(n)}(x)$ for all $x \in \mathcal{G}_S$.

Step 4 Repeat Steps 2 and 3 B times for a burn-in phase and R times for replication.

Step 5 The final density estimate $\widehat{f}_{rs}(x)$ is:

$$\widehat{f}_{rs}(x) = \frac{1}{R} \sum_{r=1}^{R} \widehat{f}_{rs}^{(B+r)}(x).$$

The algorithm for the computation of local proportions

Step 1 Initial kernel density estimation of the densities $\widehat{f}_V$ and $\widehat{f}_P$:

- Use $x_k^{(0)} = y_k$ for all $k \in U_V$ and all $k \in U_P$.
- Set the smoothing parameters $h_1^{(0)}$ and $h_2^{(0)}$ to sufficiently large values such that no spikes occur in the density estimate.
- Calculate the initial voters distribution by

$$\widehat{f}_V^{(0)}(x) = \frac{1}{N_V |H|} \sum_{k \in U_V} K\left(H^{-1}(x_k^{(0)} - x)\right).$$

- Calculate the initial party P distribution by

$$\widehat{f}_P^{(0)}(x) = \frac{1}{N_P |H|} \sum_{k \in U_P} K\left(H^{-1}(x_k^{(0)} - x)\right).$$

Step 2 Draw a stratified sample $s_V^{(n)}$ of voters and a stratified sample $s_P^{(n)}$ of party P voters.

- The strata sizes are $N_{V,a}$ for the voters and $N_{P,a}$ for the party P voters.
- The sampling of voters is with replacement from the grid $\mathcal{G}$ with sample size $N_{V,a}$ in area a. The sampling is proportional to size with $\widehat{f}_V^{(n-1)}$ as size variable. This generates $s_V^{(n)}$.
- The sampling of party P voters is with replacement from $s_V^{(n)}$ with sample size $N_{P,a}$ in area a. The sampling is proportional to size with $\widehat{f}_P^{(n-1)}$ as size variable. This generates $s_P^{(n)}$.

Step 3 Recalculate $\widehat{f}_V^{(n)}$ from the voter sample $s_V^{(n)}$ and $\widehat{f}_P^{(n)}$ from the party sample $s_P^{(n)}$.

- Determine the smoothing parameters $h_1^{(n)}$ and $h_2^{(n)}$ by the plug-in estimator of Wand and Jones (1994) from the party P sample. These smoothing parameters will be used for the estimation of both density estimates.
- Calculate $\widehat{f}_V^{(n)}(x)$ for all $x = x_{g,a}$ $(g = 1, ..., G)$ and $(a = 1, ..., A)$.
- Calculate $\widehat{f}_P^{(n)}(x)$ for all $x = x_{g,a}$ $(g = 1, ..., G)$ and $(a = 1, ..., A)$.

Step 4 Repeat Steps 2 and 3 B times for a burn-in phase and R times for replication. Compute for each replication r the ratio

$$\widehat{f}_{P|V}^{(B+r)}(x) = \frac{\widehat{f}_P^{(B+r)}(x)}{\widehat{f}_V^{(B+r)}(x)}$$

for all $x = x_{g,a}$ $(g = 1, ..., G)$ and $(a = 1, ..., A)$.

Step 5 Compute final ratio estimate $\widehat{f}_{P|V}(x)$:

$$\widehat{f}_{P|V}(x) = \frac{1}{R} \sum_{r=1}^{R} \widehat{f}_{P|V}^{(B+r)}(x)$$

for all $x = x_{g,a}$ $(g = 1, ..., G)$ and $(a = 1, ..., A)$.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Amt für Statistik Berlin-Brandenburg (2016) Wahlen zum Berliner Abgeordnentenhaus 2016. https://www.wahlen-berlin.de/wahlen/BE2016/AFSPRAES/Wahlatlas/Wahlatlas.html. Accessed 13 Jan 2020

Bradley JR, Wikle CK, Holan SH (2016) Bayesian spatial change of support for count-valued survey data with application to the American Community Survey. J Am Stat Assoc 111(514):472–487

Celeux G, Chauveau D, Diebolt J (1996) Stochastic versions of the em algorithm: an experimental study in the mixture case. J Stat Comput Simul 55(4):287–314

Groß M (2021) Kernelheaping: Kernel Density Estimation for Heaped Data. R package version 2.2.8

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. J Royal Stat Soc Ser A 180:161–183

Groß M, Kreutzmann AK, Rendtel U, Schmid T, Tzavidis N (2020) Switching between different non-herachical administrative areas via simulated geo-coordinates: a case study for student residents in Berlin. J Off Stat 36:297–314

Härdle W (1991) Applied nonparametric regression. Cambridge University Press,

Jones MC (1993) Simple boundary correction for kernel density estimation. Stat Comput 3(3):135–146

Kraak MJ, Ormeling F (2021) Cartography. Visualization of Geospatial Data, Fourth Edition edn. CRC Press,

Mitchell A (2005) Spatial Measurement and Statistics. ESRI Guide to GIS Analysis, vol 2. ESRI Press,

Oliver MA, Webster R (2015) Basic Steps in Geostatistics: The Variogram and Kriging. In Kriging (ed) SpringerBriefs in Agriculture

Rendtel U, Ruhanen M (2018) Die Konstruktion von öffentlichen Dienstleistungskarten mit Open Data am Beispiel des lokalen Bedarfs an Kinderbetreuung in Berlin. AStA Wirtsch Sozialstat Arch 12:271–284 ((the construction of public service maps with open data. the example of local need of child care in Berlin) in German.)

Rendtel U, Neudecker A, Fuchs L (2021) Ein neues Web-basiertes Verfahren zur Darstellung der Corona-Inzidenzen in Raum und Zeit. AStA Wirtsch Sozialstat Arch 15:93–106 ((a new web-based procedure for the display of corona incidences in space and time.) in German)

Tagesspiegel Wahl-Spezial (2017) Wie die Hauptstadt tickt. https://wahl.tagesspiegel.de/2017/karten/wahlbezirke/. Accessed 13 Jan 2020

Wand M, Jones M (1994) Multivariate plug-in bandwidth selection. Comput Stat 9(2):97–116