Freie Universität Berlin

Simulated geo-coordinates as a tool for map-based regional analysis

Marcus Groß
Ulrich Rendtel
Timo Schmid
Hartmut Bömermann
Kerstin Erfurth

School of Business & Economics
Discussion Paper

Economics

2018/3

Simulated geo-coordinates as a tool for map-based regional analysis

Marcus Groß[*], Ulrich Rendtel[†], Timo Schmid[‡], Hartmut Bömermann[§], Kerstin Erfurth[¶]

February 14, 2018

Abstract

Map-based regional analysis is interested to detect areas with a large concentration of certain populations. Here kernel density estimates (KDE) offer advantages over classical choropleth maps. However, kernel density estimation needs exact geo-coordinates. In a recent paper Groß et al. (2017) have proposed a measurement error model which uses local aggregates for kernel density estimation. Their algorithm simulates "exact" geo-coordinates which reflect the information on the aggregates.

In this article we suggest two extensions of this approach. First, we consider boundary constraints, which are usually ignored in the KDE framework. This concerns not only the outer limits of a municipality but also unsettled regions within a city like parks, lakes and industrial areas. Without a boundary correction standard KDEs underestimate the density in the vicinity of boundaries. Here we propose a modification of the original algorithm which uses rescaled kernel functions.

Regional maps often display local percentages, for example, voters for a special party among all voters in each voting district. Here we derive a smooth representation of percentages which is based on the ratio of two densities. Again, the original algorithm is modified to cope with the estimation of a ratio of two densities.

Our empirical examples refer to voting results from Berlin. It is shown that the proposed methodology reveals a lot of regional insight which is not produced by standard choropleth maps.

[*]INWT Statistics
[†]FU Berlin, FB Wirtschaftswissenschaft
[‡]FU Berlin, FB Wirtschaftswissenschaft
[§]Amt für Statistik Berlin/Brandenburg
[¶]Studiengang European Master of Statistics (EMOS)

Keywords: Regional Analysis, Choropleths, Kernel Density Estimation, Geo-Coordinates, Open data

1 Introduction

Regional analysis is often based on maps. The purpose of such maps is to display regional concentrations of certain populations, say, of migrants or voters for a special political party. The level of the analysis may vary: at the administrative level from NUTS 1, at the federal state level, down to NUTS 3, at the municipality level. However, often we may seek to analyse the regional distribution within the limits of a municipality, say at ZIP-code level or at the level of a voting district.

Regional analysis at this low level is confronted with several problems. First, the standard maps use choropleths, where the regional units are uniformly colored and the color pattern is restricted to a limited number, often as low as 4 or 5 levels. These choropleths incur information losses with respect to the regional position of the units inside the displayed areas and also with respect to the number of units which live in the area as their frequency is recoded to intervals that correspond to the colors of the map. With these maps it may be difficult to identify local hot spots which cross the area scheme of the map.

As a rule, the regional units are not of the same area size. There may be small areas and also large areas. Therefore their interpretation via the area sizes, which is the most appealing interpretation, can lead to wrong conclusions. As an example we display in Figure 1 the classical representation of voting results by a choropleth map. Here the voting districts, which are of different size, are the area units. The map displays the number of voters for a right wing party (AfD) in the last regional Berlin elections (2016), which is grouped into eight intervals. The general impression from Figure 1 is that the AfD is very strong in the south east of Berlin. However, as we will see, this impression is misleading. With regular regional systems of the same size, like grids of a fixed size, the above obstacle can be removed. But the visual impression of a grid map can vary substantially, if the coordinates of the grid are modified. What remains, however, is the discreteness of the representation by a discrete color scheme.

With low regional levels confidentiality issues come into play. Access to exact geo-coordinates is regarded as too risky to protect anonymity. For example, the 1 km grid maps of the German census atlas displays about one third of grid cells with grey color, which means "unsettled or to be kept secret", see `https://atlas.zensus2011.de/`. Usually, grid cells with case numbers up to three are classified as "to be kept secret". As an example Figure 2 displays the population counts in a 1 km grid of the German census in 2011 for Berlin and its vicinity. Note the large percentage of grey grid fields. Also the discontinuous representation makes it difficult to get an impression of the population density of the population in this area.

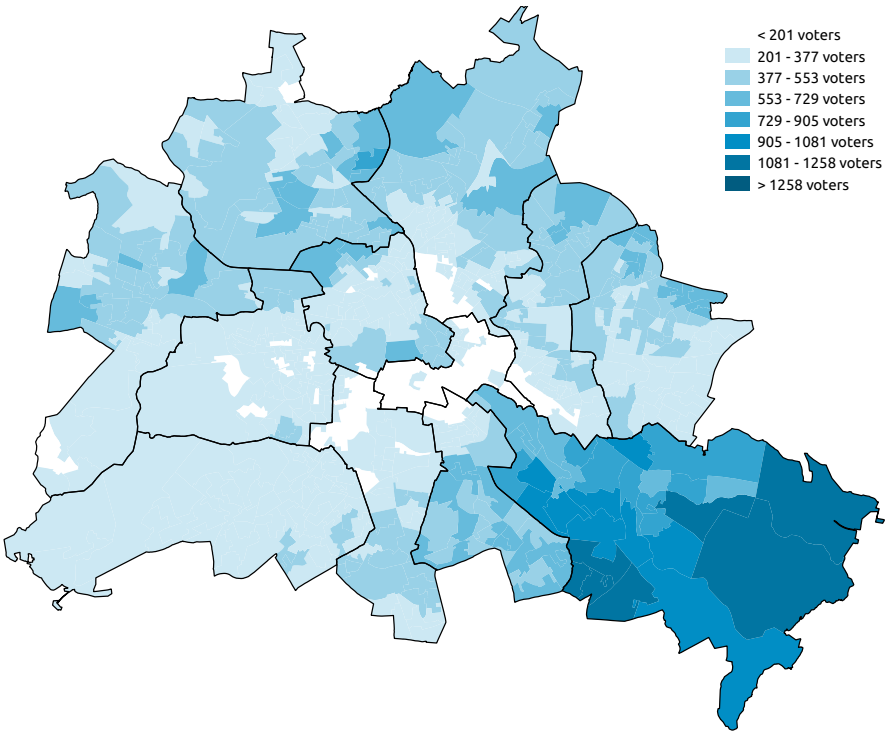

Figure 1: Choropleth map of votes for party AfD in regional elections 2016 in Berlin. Areas = Voting districts

Because of confidentiality reasons access is granted only to aggregates of larger regional units, for example, neighborhoods or, even larger, the entire municipality. Sometimes such regional information can be accessed as "Open Data" from the internet. For example, the Statistical Office of Berlin (AfS) gives access to many demographic variables at the lowest urban planning units, the so-called "Lebensweltlich orientierte Planungsräume (LOR)", see `https://daten.berlin.de/datensaetze`. However, these small regional units are of quite different size. So choropleths based on these units are far from being representative for areas.

Maps based on two-dimensional kernel densities avoid the shortcomings of choropleths as they are representative for areas and as they allow to construct highest density areas independent from administrative districts. This is an attractive tool to identify hot spots of a population of interest and/or to identify their stability over time.

However, the prerequisite of a kernel density estimate, is the knowledge of the exact geo-coordinates, which are not known in the case of regional aggregates. If the regional units are not too large one may take the centroid of the regional unit as a rough guess of the true geo-coordinate of the units in that region. Recently, Groß et al. (2017) built a bridge between the construction of density estimates and data access to regional aggregates via a measurement-error model. In their approach the true position is taken as the centroid position plus some error. The proposed algorithm starts with an initial kernel density estimate assuming all units at their centroid. From

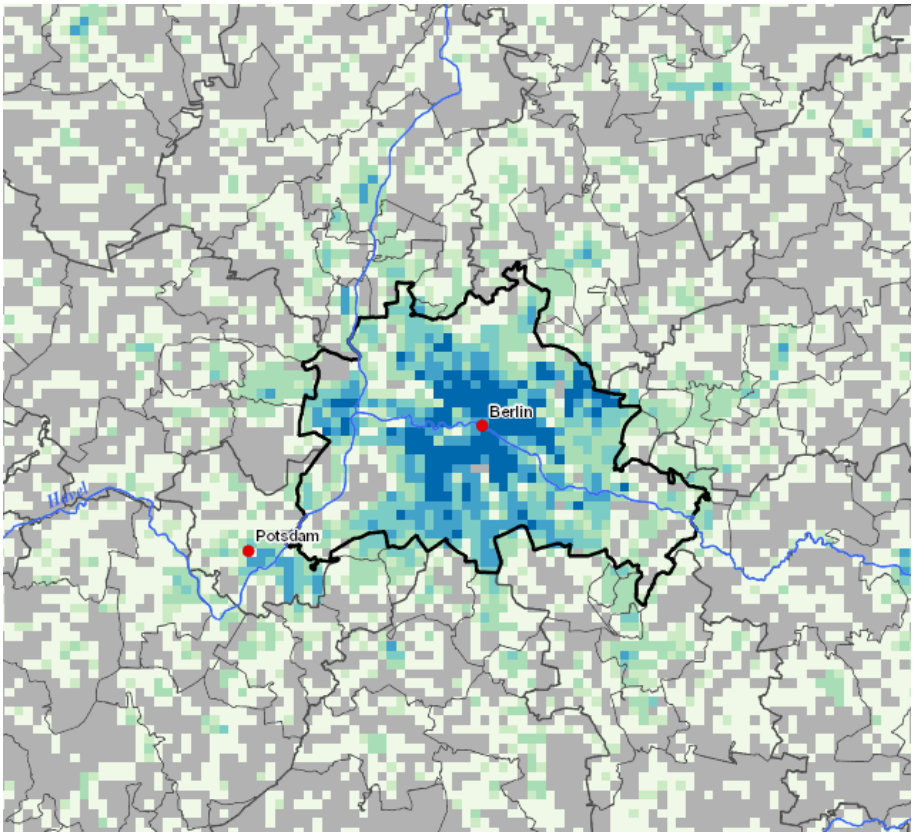

Figure 2: 1 km grid map of German census (2011) of inhabitants. Detail from Berlin and surrounding area

this estimate new coordinates are sampled by a stratified sample where the strata sizes reflect the totals of the regional units. This step gives a new set of geo-coordinates, which again are used to estimate a new density. These steps are repeated until convergence is achieved. The algorithm is an application of the Stochastic EM algorithm of Celeux et al. (1996), where the stochastic part is represented by the stratified sampling from the kernel density, and the kernel density estimation on the basis of the simulated geo-coordinates reflects the M-Step.

However, the kernel density estimates have their own methodological difficulties. Besides the fixing of the smoothing parameter, it is the inherent overlap of the resulting density over the borders of the region of interest. Even more interesting, when one is interested in low-level regional analysis, is the ignorance of the density approach about areas within the municipality that are not settled. These areas are parks, lakes, forests and industrial zones. Their proportion of the entire municipality area can be considerably. In case of Berlin these non-settled areas amount to 25 percent of the total municipality area, see Figure 3. If these areas are ignored the corresponding kernel densities are biased near the borderlines towards lower population counts. Note, that problem of external and internal borders does not occur with the choropleth approach, at least theoretically. So, for example, one may exempt the unsettled regions from the voting districts as demonstrated in Figure 3, where the unsettled regions appear grey striped. However, in every-day-use of maps larger parks, forests, lakes and industrial regions are almost never exempted from choropleths.

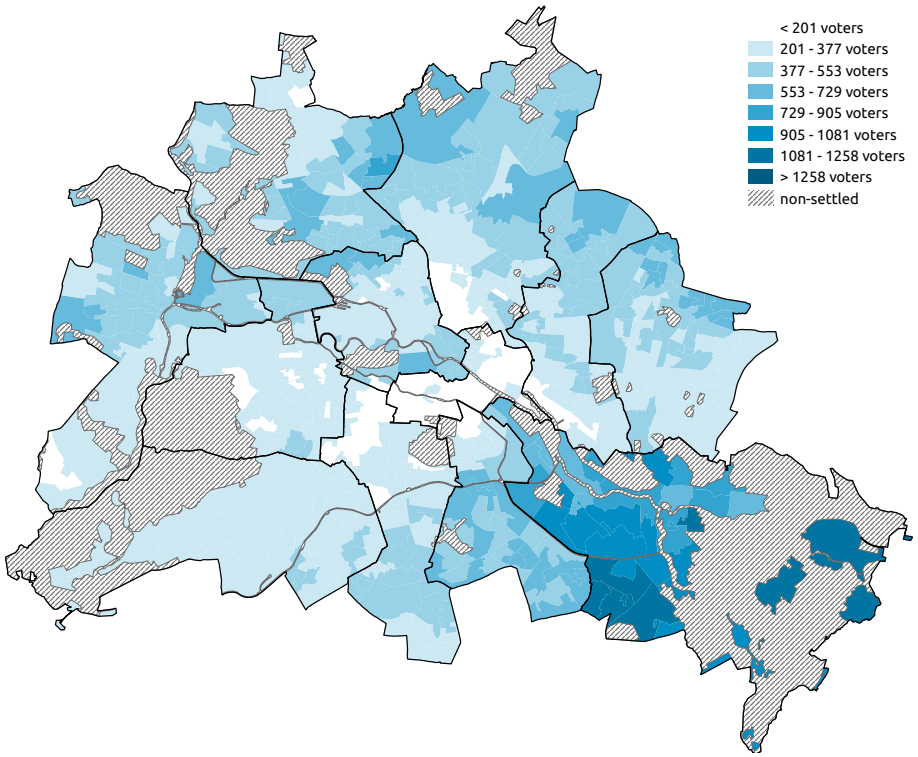

Figure 3: The display of AFD results in voting districts restricted settled areas of Berlin

While the borderline bias of the kernel density approach has been carefully investigated in one dimensional settings, less attention has been given to the two-dimensional case, see Thies (2016). One straight-forward approach would be the rescale the kernel functions at each point to have a volume of 1 over the settled area. This approach was evaluated in a simulation study for Berlin for exact geo-coordinates, see Thies (2016). Here we will adapt this approach to the SEM algorithm.

Standard choropleths do not only display absolute figures related to one area. Often they display percentages, for example, percentages of voters for a special party within a voting district. On the first sight it is not obvious how percentages can be embedded into the framework of densities. Here we will demonstrate how the ratio of densities, namely the density of the party P voters and the density of all voters, can be used to derive a local percentage at the level of the pixels. We also demonstrate the use of such intensity maps to display areas with high local concentration of voters of a special party.

The article is organized as follows: Section 2 resembles the general SEM algorithm. Section 3 discusses the necessary modification to deal with the borderline problem. Section 4 deals with the treatment of percentages. It displays the necessary modification of the SEM algorithm for the computation of regional percentages. Section 5 is devoted to the application of these techniques to display the voting results for the election of the Berlin regional parliament in 2016. Section 6 concludes.

2 The general algorithm for simulated geo-coordinates

Let the areas be indexed by $a = 1, \ldots, A$. For each area the total $N_a \quad (a = 1, \ldots, A)$ of the variable of interest is known. Then the total population U is of size $N = \sum_{a=1}^{A} N_a$. U may be divided into A strata U_a with stratum size N_a. For $k \in U_a$ we assume y_k, which is the coordinate of the centroid of area a, to be a reasonable approximation of the true geo-coordinate of x_k of unit k.

Now the standard kernel density estimator $\hat{f}(x)$ of the population density $f(x)$ of the variable of interest at geo-coordinate x is

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

where K is the kernel function and H is a two-dimensional smoothing matrix. Here we will use $H = diag(h_1, h_2)$, with suitably chosen smoothing parameters h_1 and h_2. We use the plug-in approach of Wand and Jones (1994) for bandwidth selection. For our analyses we use the Gaussian Kernel function $K(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x'x)$.

To keep things numerically tractable we will generate X-coordinates only on a fine grid of geo-coordinates. Also, we will evaluate the resulting density estimate only on these grid-points. Let $x_g \quad (g = 1, \ldots, G)$ be the geo-coordinate of the G grid points. Then the set $\mathcal{G} = \{x_g | g = 1, \ldots, G\}$

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.

The basic SEM algorithm may be formulated as follows:

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

- Use $x_k^{(0)} = y_k$ for all $k \in U$
- Determine smoothing parameters $h_1^{(0)}$ and $h_2^{(0)}$
- Calculate $\hat{f}^{(0)}(x)$ for all $x = x_{g,a}$ for all $g = 1, \ldots, G$ and all $a = 1, \ldots, A$

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

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

Step 3 Recalculate $\hat{f}^{(n)}$ from sample $s^{(n)}$

- Determine the smoothing parameters $h_1^{(n)}$ and $h_2^{(n)}$.
- Calculate $\hat{f}^{(n)}(x)$ for all $x = x_{g,a}$ for all $g = 1, \ldots, G$ and all $a = 1, \ldots, A$

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 $\hat{f}(x)$ is:

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

This algorithm can be realized with the *R*-package *kernelheaping*, see Groß (2016).

3 The boundary correction of the kernel estimate

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:

$$w_x = \int_{\mathcal{S}} \frac{1}{|H|} K(H^{-1}(x - y)) dy \tag{2}$$

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

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

$$\hat{f}_{rs}(x) = \frac{1}{N|H|} \sum_{k \in U} \frac{1}{w_x} K(H^{-1}(x - x_k)) \tag{3}$$

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

$$w_x \approx \sum_{y \in \mathcal{G}_S} \frac{1}{|H|} K(H^{-1}(x - y)) \Delta_{\mathcal{G}} \tag{4}$$

In the case of a Gaussian Kernel we obtain:

$$w_x = \frac{\Delta_{\mathcal{G}}}{\sqrt{2\pi}} \frac{1}{h_1 h_2} \sum_{(y_1, y_2) \in \mathcal{G}_S} \exp\{-0.5(\frac{(x_1 - y_1)^2}{h_1} + \frac{(x_2 - y_2)^2}{h_2})\} \tag{5}$$

w_x is to be 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 w_x have to be recalculated for every change of H. This can happen in every round of the modified SEM algorithm displayed below.

Now the modified SEM algorithm below computes the rescaled kernel density estimate $\hat{f}_{rs}$:

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

- Use $x_k^{(0)} = y_k$ for all $k \in U$: All units are supposed to lay in 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.
- Determine smoothing parameters $h_1^{(0)}$ and $h_2^{(0)}$.
- Compute weights $w_x^{(0)}$ for every $x \in \mathcal{G}_S$.
- Calculate $\hat{f}_{rs}^{(0)}(x)$ for all $x \in \mathcal{G}_S$.

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

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

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

- Determine the smoothing parameters $h_1^{(n)}$ and $h_2^{(n)}$

- Determine adapted weights $w_x^{(n)}$ for every $x \in \mathcal{G}_S$
- Calculate $\hat{f}_{rs}^{(n)}(x)$ for all $x \in \mathcal{G}_S$

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

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

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

4 The estimation of proportions

Let f_V be the two dimensional density of voters. Correspondingly let f_P be the two dimensional density of voters of party P. The naming refers to the application in voting analysis. However, P can be any variable which creates a subset of the universe of voters. 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 $N_V f_V(x_1, x_2)\Delta_{x_1} \times \Delta_{x_2}$. Similarly, the expected number of voters for party P at coordinate $x = (x_1, x_2)'$ is obtained by $N_P f_P(x_1, x_2)\Delta_{x_1} \times \Delta_{x_2}$. Hence the ratio $f_{P|V}(x_1, x_2) = \frac{N_P}{N_V} f_P(x_1, x_2)/f_V(x_1, x_2)$ 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.

The estimation of $f_{P|V}(x) = f_{P|V}(x_1, x_2)$ can be done straightforward with the basic SEM algorithm or its boundary corrected version of the previous sections. However, the estimation of f_P should not be carried out independently from the simulated geo-coordinates that were generated for the voters. Instead, as all voters for party P are voters, their distribution should be concentrated on the coordinates of the voters. If $s_V \subset \mathcal{G}$ denotes the sample of grid points selected for the voters, then for logical consistency the sample s_P of voters for party P should be a subset of s_V.

Let U_V be the universe of voters. For each area a the number of voters is $N_{V,a}$. The total number of voters is N_V. Similarly we obtain for party P voters U_P, $N_{P,a}$ and N_P.

Note, that for fixed geo-coordinates and equal smoothing factors for voters and party P voters $\hat{f}_{P|V}$ is algebraically equivalent to the Nadaraya-Watson estimator. To see the equivalence, 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 $\hat{f}_{NW}$ is then given by (Härdle (1991)):

$$\begin{aligned} \hat{f}_{NW}(x) &= \frac{\frac{1}{N_V} \sum_{k \in U_V} \frac{1}{|H|} K(H^{-1}(x - X_k)) P_k}{\frac{1}{N_V} \sum_{k \in U_V} \frac{1}{|H|} K(H^{-1}(x - X_k)} \\ &= \frac{N_P \hat{f}_P(x)}{N_V \hat{f}_V(x)} \end{aligned} \tag{6}$$

Therefore, we obtain the following algorithm:

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

- Use $x_k^{(0)} = y_k$ for all $k \in U_V$ and all $k \in U_P$
- Determine smoothing parameters $h_1^{(0)}$ and $h_2^{(0)}$
- Calculate the initial voters distribution by

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

- Calculate the initial party P distribution by

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

Step 2b 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 $\hat{f}_V^{(n-1)}$ 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 $\hat{f}_P^{(n-1)}$ as size variable. This generates $s_P^{(n)}$.

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

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

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

Compute for each replication r the ratio

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

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

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

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

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

5 Application to voting results of the Berlin regional parliament

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. The overall percentage for the AfD was 14.1 %.

The densities for the distribution of voters are normalized to a volume of 1 under their surface. In order to make them comparable they shoud be multiplied by the absolute number N_P of voters for party P. If we multiply the densities with the area of the pixels, which is 141×141 m^2 in our case, we end-up with a scale which can be interpreted as the number of party P voters per pixel.

Figure 4 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.

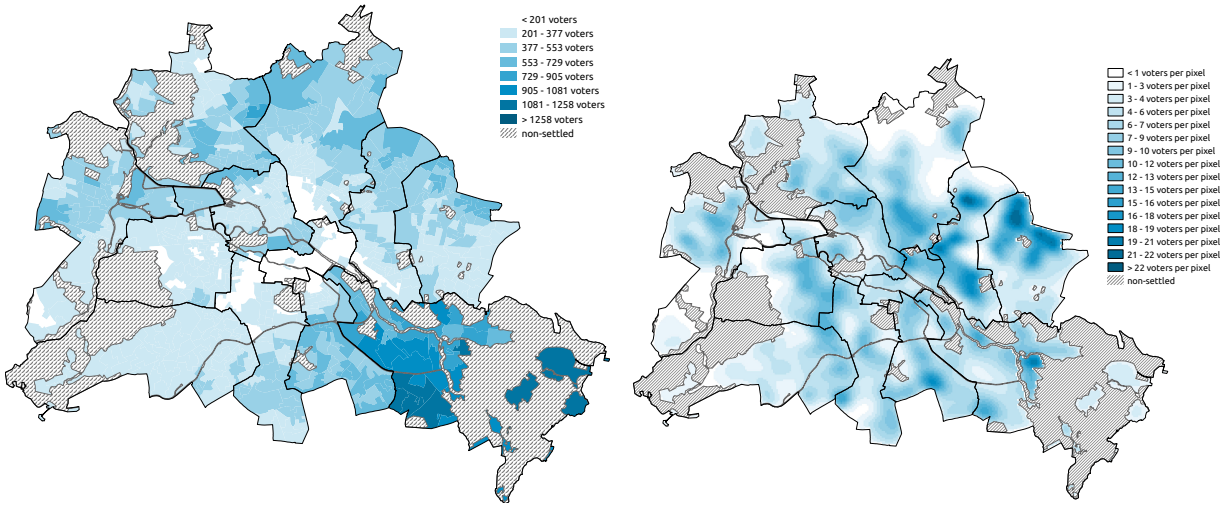

Figure 4: Absolute number of voters for party AfD in regional elections 2016 in Berlin. Left: display by Choropeth, areas =voting districts Right: number of voters per pixel

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.

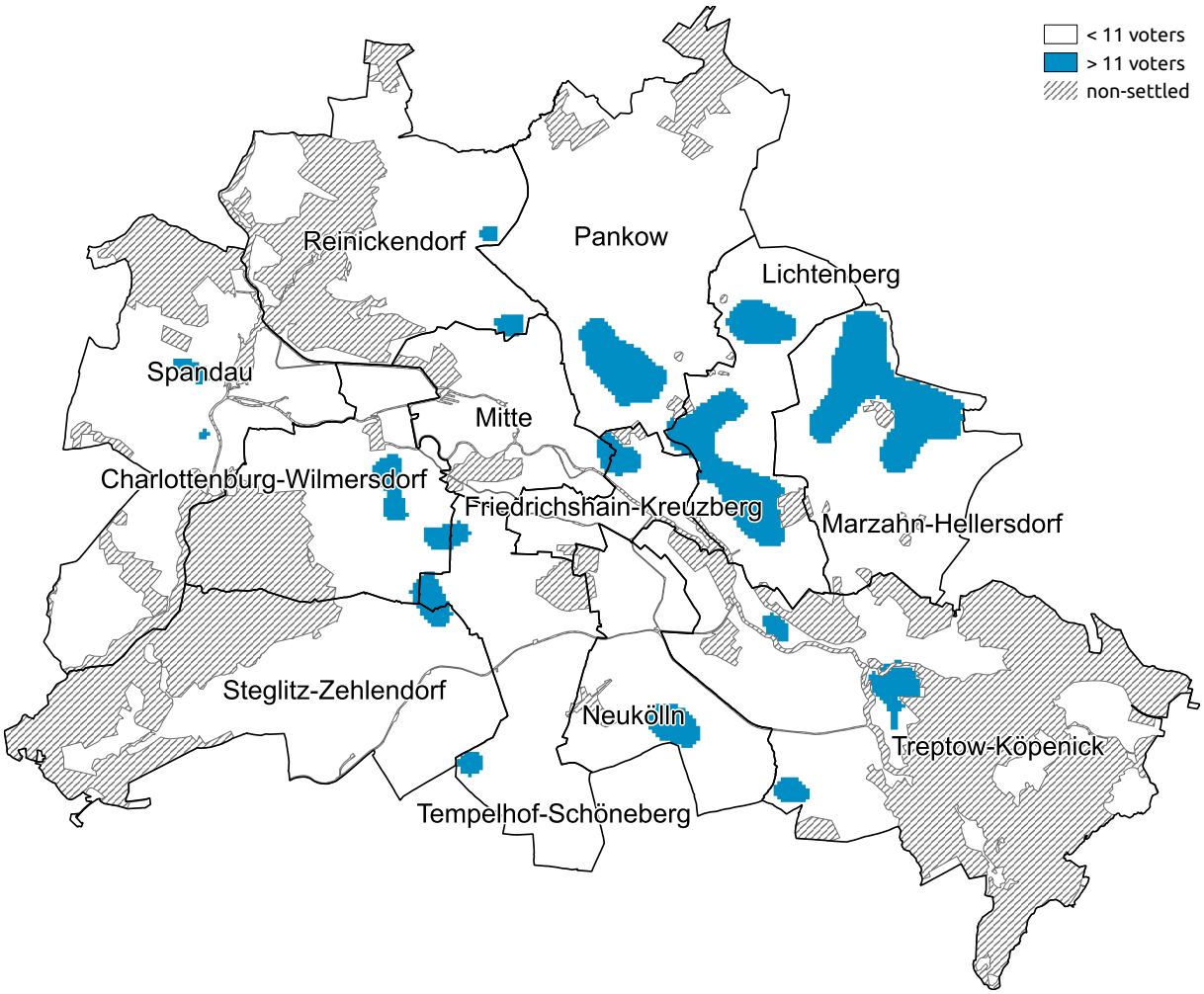

Figure 5: High density area covering 20 percent of AFD voters

Figure 5 displays the high density area for AfD voters. The displayed area covers 20 % of all AfD voters. 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 quarters with tower building flats from the 70-s to the 90-s of the last century. This does not only hold for former East-German settlements in the district Mahrzahn-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 Figure 4

A different attractive feature is the comparability of the re-scaled densities for different parties.

So one can display for each area the party which achieves the highest number of voters per pixel. Figure 6 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).

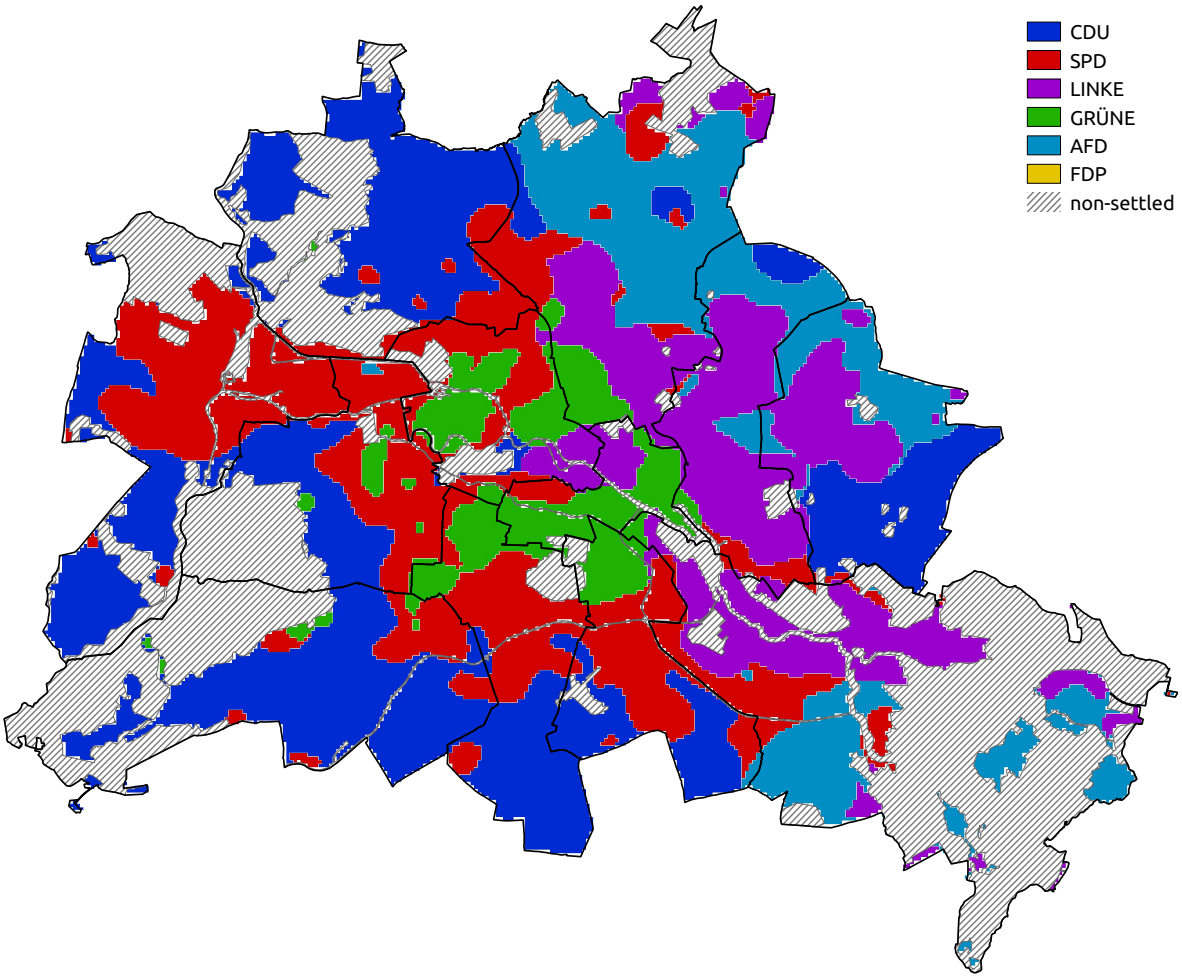

Figure 6: The winner with respect to the highest number of voters per pixel. Legend: CDU= dark blue, SPD =red, Linke= purple, Grüne =green, AFD = light blue

If we switch to the estimation of local percentages we first have to estimate the distribution of the voters. Figure 7 displays a density estimate of the distribution of voters per pixel. This density varies considerably within Berlin which is the reason why the choropleth maps of absolute figures are so misleading in this case.

Figure 8 compares the local proportions of AfD voters via density estimation with the results from voting districts. There is a high coincidence of results in the two maps, displaying high percentage numbers in the south-east and the north-east of Berlin. However, the percentages of the single voting districts are somewhat erratic and don't offer a combination of adjacent voting districts to low and high percentage areas.

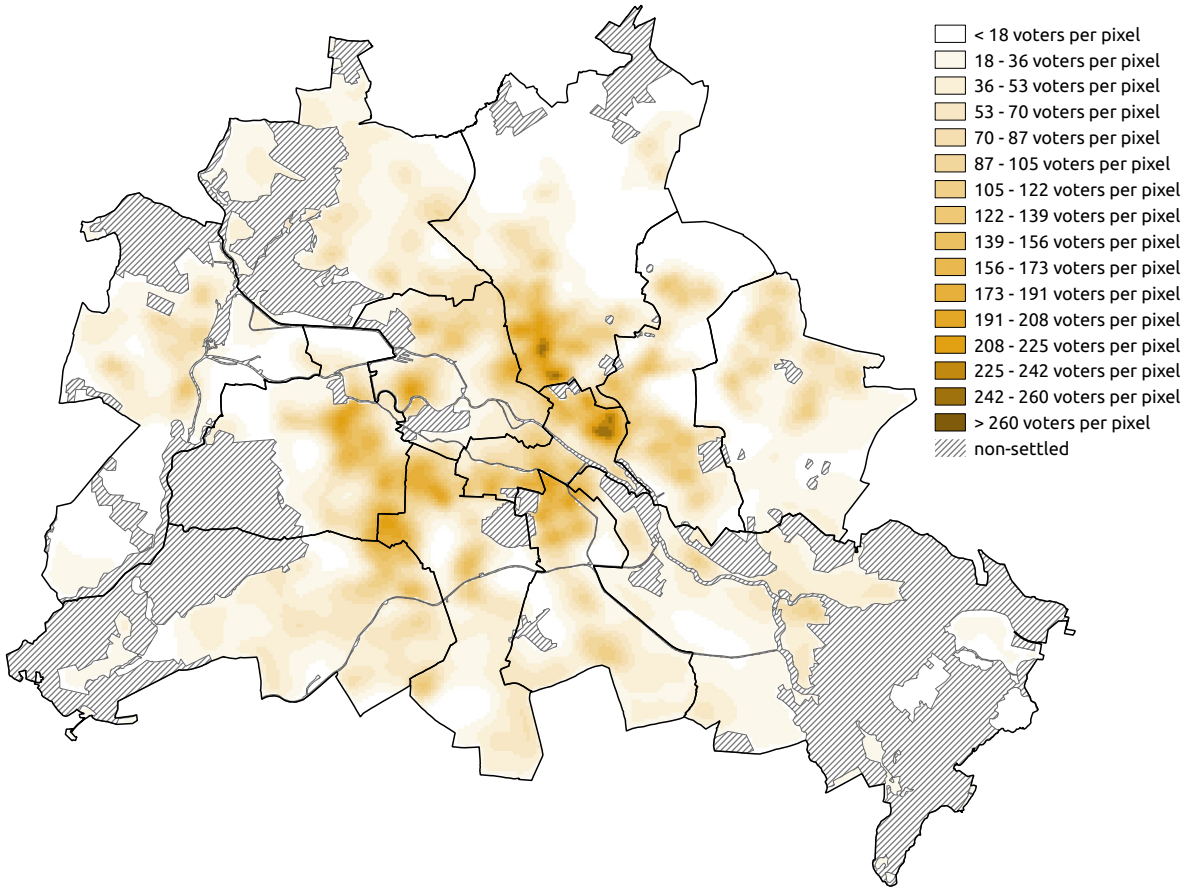

Figure 7: The number of voters per pixel

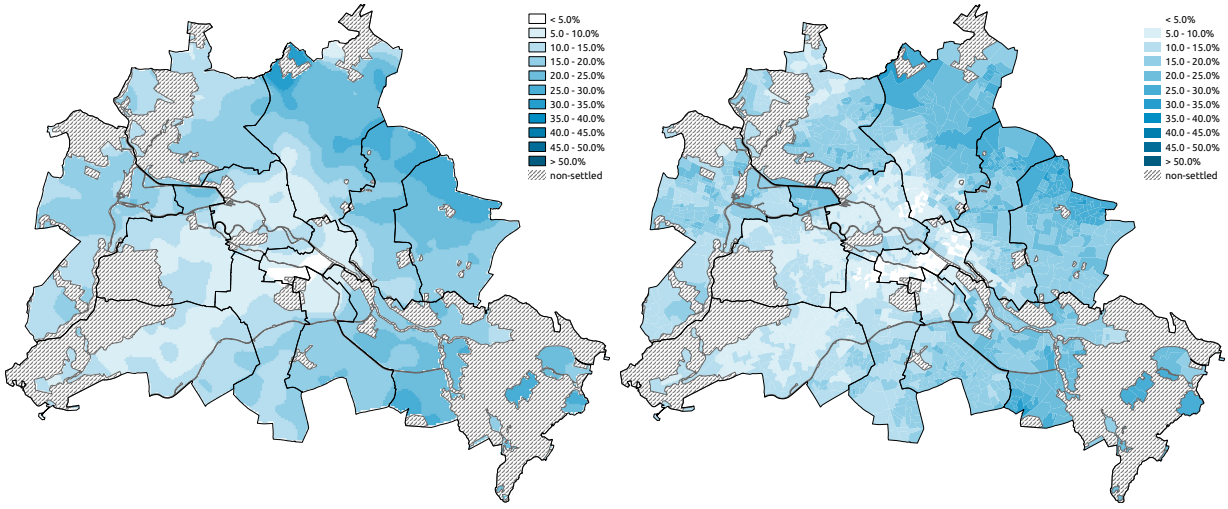

Figure 8: Percentage of AFD-Voters: Left: Local proportions via densities, Right: Proportions in voting districts

This is the great advantage of the density approach. Here it is possible to create smooth high percentage areas. There are two versions of such high percentage areas. The first version asks for the area where a prefixed limit is exceeded. Such an area is shown in Figure 9 for a limit of 10 percent. It displays for broad regions a substantial support of the AfD.

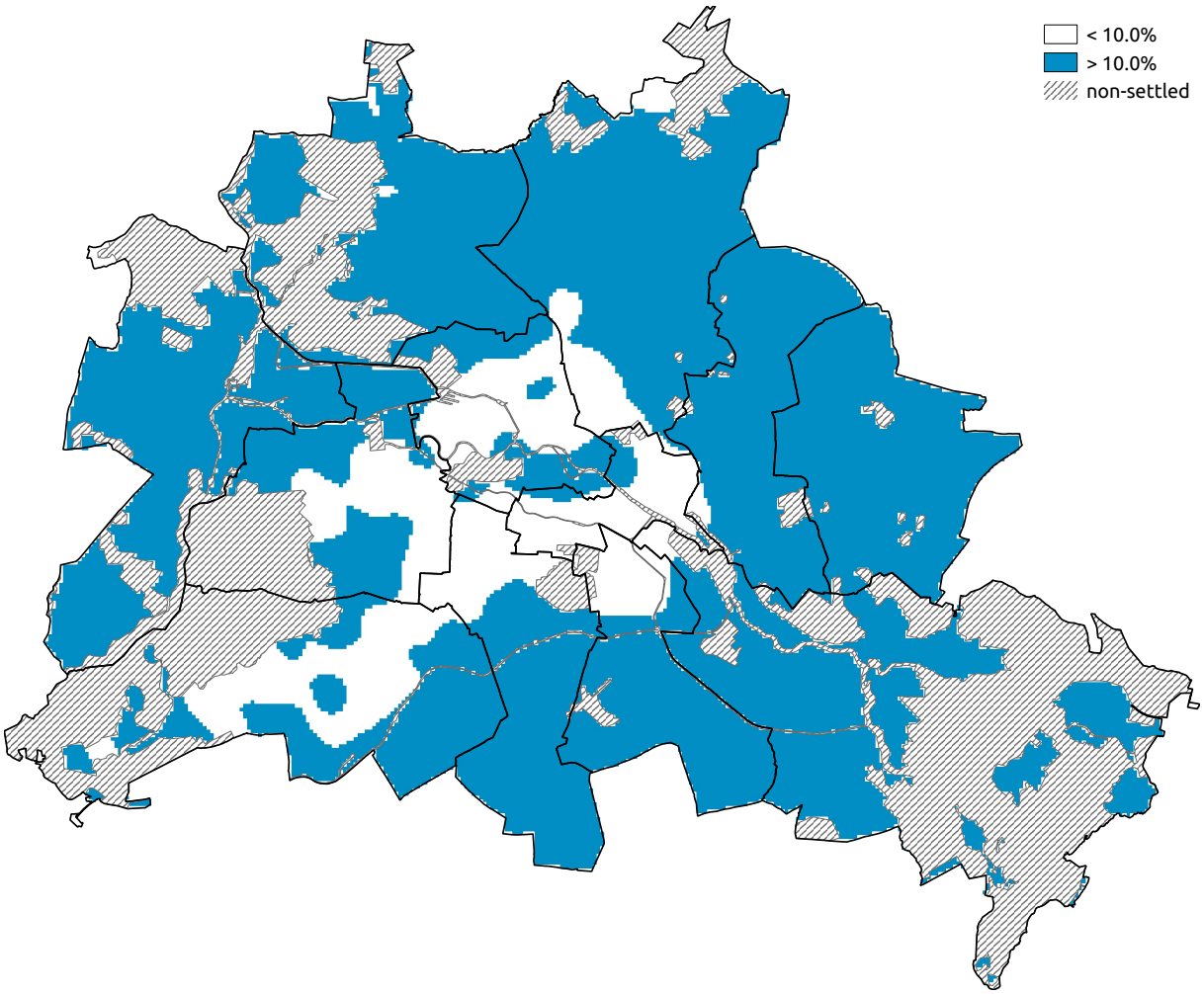

Figure 9: High percentage areas: Percentage for AfD is larger than 10 %

The second possibility to display high percentage areas is to keep the percentage of the covered area fixed, say 20 percent, and to ask for the limit which defines the borderline of this area. Such a display is convenient for comparisons between different parties. Figure 10 compares the high percentage areas for the six parties which became elected into the parliament. For each party the 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 party

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

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

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 the party with the highest percentage. Because of the smooth shape of the local percentages their maximum is also smooth. Figure 11 compares a map of the local winners derived from the densities with a choropleth which displays for each voting district the color of the party with the maximum percentage in the district. Despite the different construction the two maps give a similar impression where the respective parties have a local majority.

6 Concluding Remarks

The basic idea of the density approach presented here is to produce maps with smooth concentration areas. The rationale of this idea in our examples is that party preferences do not vary in a discontinuous manner, like the choropleths suggest. However, the extension of the density approach to respect unsettled areas and the borderlines of the city introduces some kind of discontinuity. The degree of smoothing depends on the number of observations, which is given in our example by the number of voters in Berlin, which is about $N_V = 1.6$ million voters. Because of these very high case numbers also small regional differences were well displayed in our examples. In the case of party percentages there was a very good coincidence with the choropleth map but without their intrinsic discontinuities.

An alternative approach with respect to smoothing is the interpolation by spline approximation, see for example Fahrmeir et al. (2013). Here, we could interpolate the proportion of voters of a certain party at the centroids of the voting districts. With this approach the degree of smoothing is regulated by technical parameters, like the degree of the interpolating polynom, the number of knots or the size of a penalizing term in case of penalized splines. Furthermore, the resulting map of interpolated party proportions lacks the intuitive interpretation as a local limit of a proportion of a certain party. Also smoothing via splines does not use the geographical form of the voting districts, which is used in our approach.

It is the aim of a regional analysis to link information on local concentrations with regional

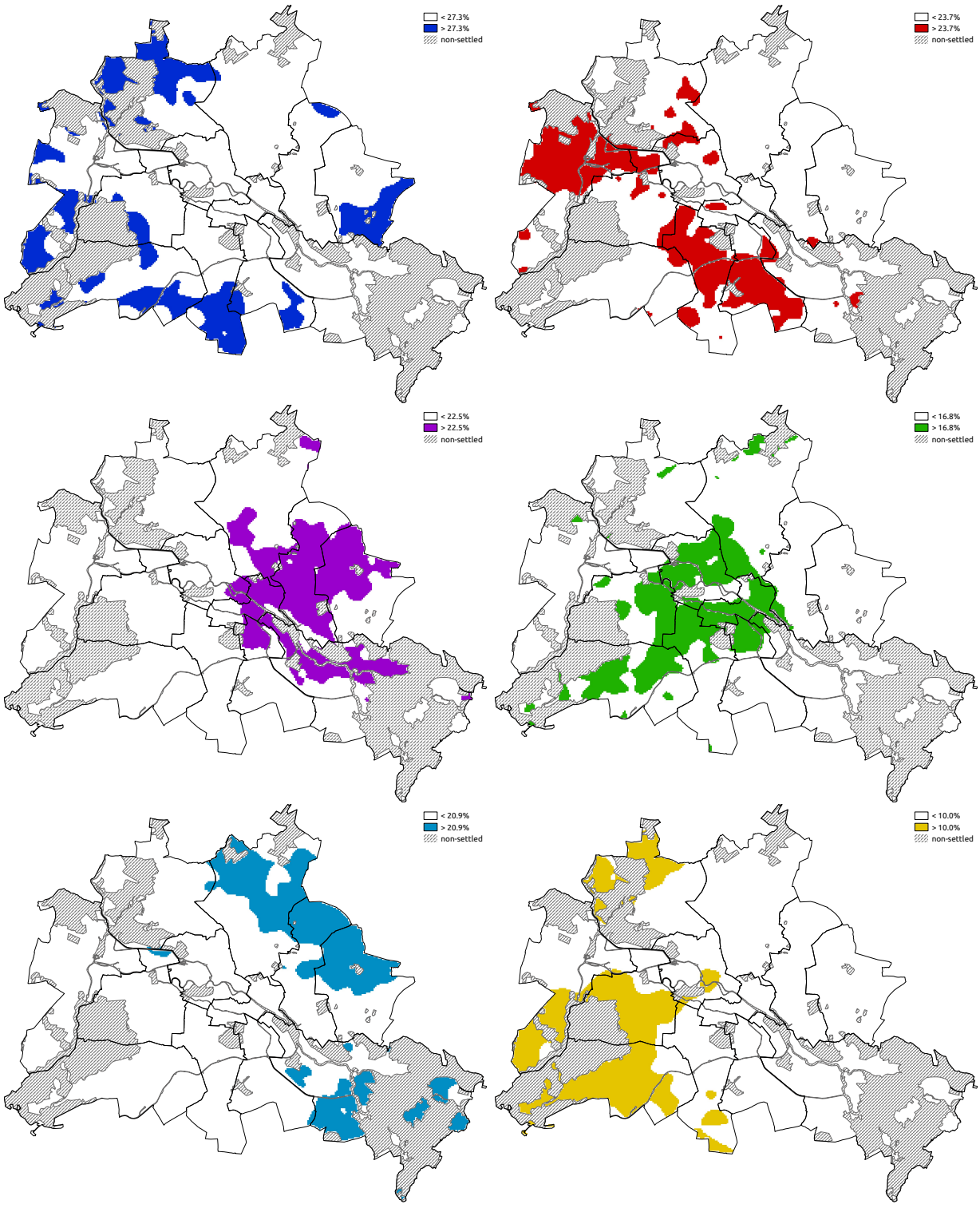

Figure 10: High percentage areas for 6 parties: Top: Left: CDU (dark blue), Right: SPD (red); Middle: Left: Linke (purple) , Right: Grüne (green), Bottom: Left: AfD (light blue), Right: FDP (yellow). Covered area is 20 % of the settled population

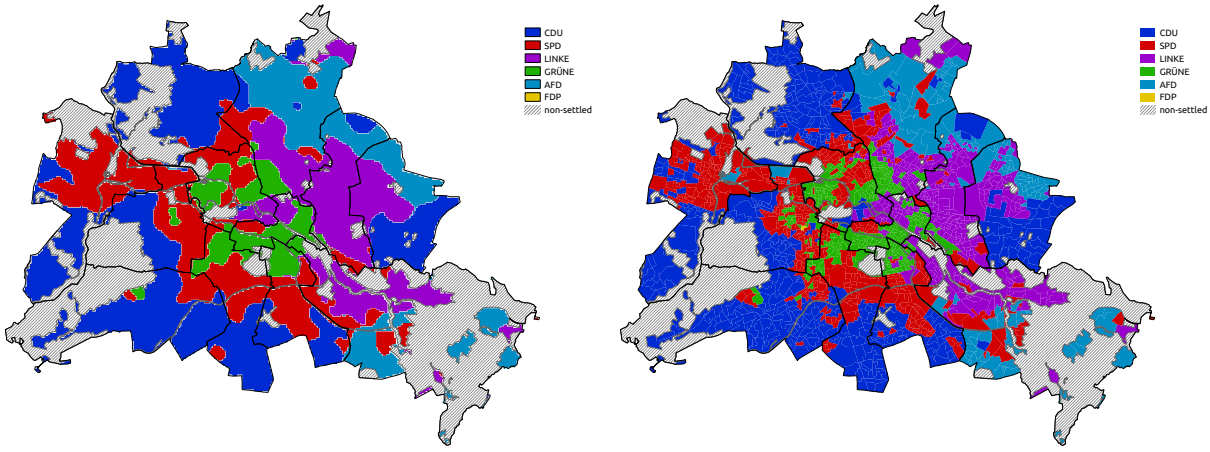

Figure 11: The party with the highest local percentage compared to the winner of the voting districts: CDU (dark blue), SPD (red), Linke (purple) , Grüne (green), AFD (light blue), FDP (yellow).

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. In this context kernel density estimates are often referred to as "heat maps". However, it is generally assumed here that the geo-coordinates are exactly known.

With the approach presented here it is possible to produce density maps not only for voting variables but for all variables with a discrete measurement of regional totals. Often these variables and the information on their local totals can be accessed via an open data portal; for example, the open data portal of Berlin may be reached by `https://daten.berlin.de/`. In Figures 12 and 13 the local aggregates from 447 local planning unit districts on children were used to estimate their regional density in Berlin. Then the geo-coordinates of local service units are displayed as dots in the map. From such a service map it is easy to identify mis-allocations of service units, see Ruhanen (2018). For example, Figure 12 displays the density of children under the age of 6. Here the dots display the location of Kindergardens. The figure clearly demonstrates that the strong local concentrations of children are well reflected by the allocation of kindergardens. Contrarily, the allocation of pediatrists in Figure 13 does apparently not meet the concentration areas of children below the age of 18.

Therefore our approach opens a lane to a broad and efficient use of publicly available micro data for the planning and formulation of service needs.

One particular feature of the density approach is the possibility to compare the evolution of high density areas over time. It is attractive for the regional analysis of voting results to see whether the high density regions are stable over time or not, see for example the analysis of the

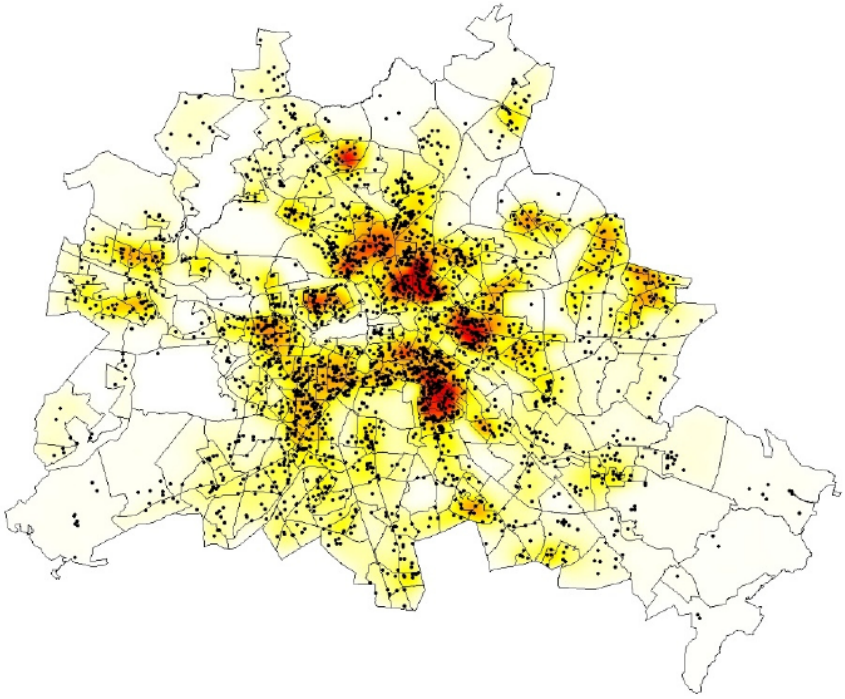

Figure 12: The estimated density of children under the age of 6 from regional aggregates and the allocation of Kindergardens in Berlin. Graphic taken from Ruhanen (2018)

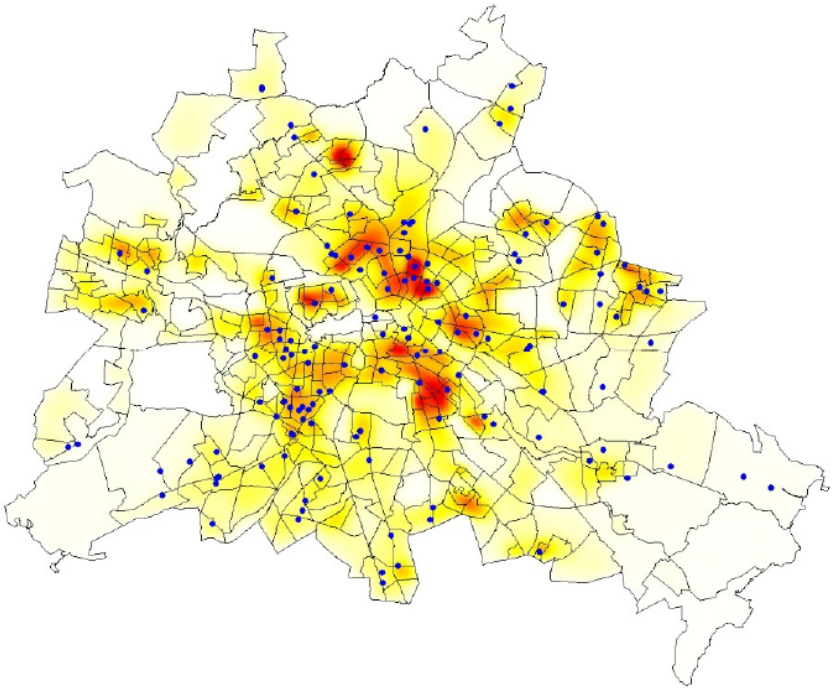

Figure 13: The estimated density of children under the age of 18 from regional aggregates and the allocation of pediatrists in Berlin. Graphic taken from Ruhanen (2018)

Berlin results for the elections of the German Bundestag over seven election periods (1990 to 2013) under `https://wahl.tagesspiegel.de/2017/karten/berlin/`. With the choropleth approach one is immediately confronted with the problem of changing shapes of the voting districts for subsequent elections. Here one would have to recalculate the former voting results to the actual voting districts, which turns out to be a tedious work. Note, that this problem does not occur with our approach, as the resulting map is independent from the aggregates which were used for the computation of the density.

References

Celeux, G., D. Chauveau, and J. Diebolt (1996). Stochastic versions of the em algorithm: an experimental study in the mixture case. Journal of Statistical Computation and Simulation 55(4), 287–314.

Fahrmeir, L., T. Kneib, S. Lang, and B. Marx (2013). Regression: Models, Methods and Applications. Springer.

Groß, M. (2016). Kernelheaping: Kernel Density Estimation for Heaped Data. R package version 1.6.

Groß, M., U. Rendtel, T. Schmid, S. Schmon, and N. Tzavidis (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 of the Royal Statistical Society: Series A (Statistics in Society) 180, 161 – 183.

Härdle, W. (1991). Applied Nonparametric Regression. Cambridge University Press.

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

Ruhanen, M. (2018). Die konstruktion von öffentlichen dienstleistungskarten mit open data am beispiel des lokalen bedarfs an kinderbetreuung in berlin. (the construction of public service maps with open data. the example of local need of child care in berlin) in german. Thesis at Economic Department of Freie Universität Berlin.

Thies, S. (2016). Boundary correction for two-dimensional kernel density estimation. Thesis at Economic Department of Freie Universität Berlin.

Wand, M. and M. Jones (1994). Multivariate plug-in bandwidth selection. Computational Statistics 9(2), 97–116.

Diskussionsbeiträge - Fachbereich Wirtschaftswissenschaft - Freie Universität Berlin
Discussion Paper - School of Business and Economics - Freie Universität Berlin

2018 erschienen:

2018/1 BESTER, Helmut und Ouyang YAOFU
Optimal Procurement of a Credence Good under Limited Liability
Economics

2018/2 GROß, Markus, Ulrich RENDTEL, Timo SCHMID und Nikos TZAVIDIS
Switching between different area systems via simulated geo-coordinates: a case study for student residents in Berlin
Economics