Rendtel, Ulrich; Gril, Lorena

Working Paper

Kernel density smoothing as a means for the construction of anonymized regional maps

Discussion Paper, No. 2025/6

Provided in Cooperation with:
Free University Berlin, School of Business & Economics

Suggested Citation: Rendtel, Ulrich; Gril, Lorena (2025) : Kernel density smoothing as a means for the construction of anonymized regional maps, Discussion Paper, No. 2025/6, Freie Universität Berlin, School of Business & Economics, Berlin, https://doi.org/10.17169/refubium-48093

This Version is available at:
https://hdl.handle.net/10419/323243

FREIE UNIVERSITÄT BERLIN

Kernel density smoothing as a means for the construction of anonymized regional maps

Ulrich Rendtel
Lorena Gril

School of Business & Economics
Discussion Paper

Economics

2025/6

Kernel density smoothing as a means for the construction of anonymized regional maps.

Ulrich Rendtel

Lorena Gril*

July 24, 2025

Abstract

The smearing effect of kernel estimates of the local density, local proportions and local means is used as a means for the construction of anonymized maps. The standard anonymization criteria were derived for the display of case numbers of a predefined area system. However, for kernel estimates there does not exist such a defined area system. We discuss the resulting difficulties of the application of these criteria for kernel estimates. Besides, there are some de-anonymization risks which are specific for kernel estimates. We discuss these topics for data from 1.9 million Berlin taxpayers with known exact address and taxable income. In the conclusions we vote for a much stronger emphasis on the output format of a map and the labelling of the displayed values in the map.[1]

Keywords: Regional maps, Kernel density estimation, Anonymity, Choropleth maps, Taxpayers.

*We thank for the support of this project, the disposal of the data base and data access by the Amt für Statistik Berlin-Brandenburg.

[1]We thank Martin Möhler (Statistisches Bundesamt, Destatis) for critical reading and constructive comments.

1 Introduction

Regional maps based on georeferenced data have become a topic of increased interest for social scientists and political consulting of targeted programs. This interest in detailed regional information is confronted with the need to protect the confidentiality of the data. Although data protection and confidentiality of individual data are a general concern the access to regional information at a small geographic scale is highly sensitive to the risk to disclose the anonymity of the individual information. The more precise the regional information is and the lower the number of persons in a specified area is the easier it is to identify a person in a survey.

The display of maps is a specific form of regional data analysis. There are two basic schemes: frequency maps, which display the number of persons with a certain characteristic in the respective areas, and mean value maps, which display the average value of a metric variable of interest in the respective areas.

A straightforward approach to establish confidentiality is to apply the privacy rules developed for tabular analysis in official statistics, see Hundepool et al. (2012) and Eurostat (2025). In regional analysis the cells of the tables are either defined by administrative entities, like counties, municipality areas or neighbourhoods, or by regular grids cells of different sizes. Cells which fail the anonymity criterion are suppressed in the map, for example, by graying them. Or, they are collapsed with neighbouring cells until they meet the anonymity criteria, see Lagonigro et al. (2021) or Haddam et al.(2020). A different approach is the so-called cell key method, which basically adds random noise to cell counts in a regularized form, see Meindl (2023). Here all cells counts below a critical value are either shifted above this limit or set to zero, depending on a pseudo-random mechanism (see, for instance, the example in Hundepool (2024)[2]) This feature prevents the suppression of the respective areas in the map, see, for example, the German Census-Atlas 2022 (Destatis 2024) as an example of the application of the cell key method (Enderle and Kleber 2024).

The kernel smoothing approach as an anonymisation strategy was proposed by de Jonge and de Wolf (2016) and de Wolf and de Jonge (2018). They use smoothing by kernel density estimates

[2] `https://sdctools.github.io/handbookSDC/05-frequency-tables.html#sec-CKM_freq`

(Silverman 1986) in the case of frequency data and the nonparametric Nadaraya-Watson regression estimator (Härdle 1991) in the case of mean value maps. Because of the smearing effect of the kernel function one may expect a high anonymization effect on the resulting maps. Meanwhile the kernel smoothing approach is available in the R-Package *sdcSpatial* (de Jonge 2022).

We regard the kernel density smoothing as a means to produce anonymized output on the basis of non-anonymized data. With respect to the output one may distinguish different types of maps:

1. The map may be in an interactive format with machine-readable values, or be a static picture (digital or print).

2. The users may or may be not allowed to zoom-in the map below a predefined level.

3. The users may have access to the exact values of the estimated kernel density, or be restricted to maps with interval labeling, say, six intervals.

4. The users may or may not have access to the exact geocoordinates of the population displayed in the maps.

5. The users may or may not have access to the information on the kernel function and the smoothing value which were used for the generation of the map.

The maps are to be distinguished from maps which are based on already anonymized geo-coordinates like area aggregates or geo-masked individual data. In this case the anonymization of the geo-coordinates may be regarded as a measurement error of the true geo-coordinates. This measurement error has to be considered in the estimation of the original density (Rendtel and Schmid 2024). In the case of regional aggregates one has to simulate geo-coordinates. These simulated coordinates are the basis of kernel smoothing routines, see Groß et al. (2017,2020) and Erfurth et al. (2022). But also in the case of geo-masked data one can improve the accuracy of a naive use of randomly disturbed geo-coordinates, see Hossain (2023). As the output bases on already anonymized data one expects no further privacy problems.

In order to demonstrate the problems and benefits which arise from the use of kernel smoothing techniques we use geo-coded data on tax payers in Berlin. Within the Amt für Statistik Berlin-Brandenburg we got access to 1.9 million tax payer records of the 2019 tax cohort including

access to their exact addresses. This data set served as a gold standard in the measurement error framework of anonymization by aggregation. Here we compared the statistical properties of the density estimates based on the anonymized data with the kernel density estimates based on the exact geocoordinates. As a by-product we could study the smoothing effect of the kernel density estimation as an anonymization tool. These results are reported here.

Before we study the anonymization by kernel smoothing we present an empirical result which displays the low capacity of the standard choropleth maps to unfold regional clusters which are easily detected by kernel smoothing. Section 3 shortly introduces three kernel smoothing estimates for (a) density estimation, (b) local proportions of a subpopulation and (c) the local mean of a metric variable. We then display the anonymization criteria which were derived in the context of tabular analysis. We report the difficulties which arise from the application of these criteria in the context of kernel smoothing maps. Section 5 demonstrates the smoothing effect of kernel estimates of regional ratios. In Section 6 we study the effect of area suppression as a function of different minimum frequency rule rules. There we also discuss a mixture of area suppression and local proportion estimates via kernel estimates. There are also de-anonymization risks which are specific for kernel density estimates. They are discussed in Section 7. These risks arise from isolated observations and the detection of individual values from access to machine-readable kernel estimates. Here we also discuss the role of the kernel smoothing parameter as an anonymization parameter. In the concluding remarks we summarize and stress the importance of the output format of the maps.

$x_{\mathbf{g}} \quad (\mathbf{g} = 1, \ldots, G)$

2 Kernel density smoothing as an alternative to Choropleth maps

The standard maps are based on area counts. In the geographic literature these maps are called choropleth maps. If the area system refers to meaningful entities, like municipalities or neighborhoods, these counts have a direct interpretation. However, if we change to regular grid data there

is no such immediate interpretation. On the contrary, the smaller the grid-size becomes, the more noisy becomes the resulting choropleth map. Here a density approach results in maps where local population clusters are easily identified.

As example we use a 10 percent simple random subsample of the Berlin taxpayers which has still 190 thousand observations. The reason for subsampling was twofold: first to reduce the computational burden and second to preserve the anonymity of the data. Figure 1a displays the count data of taxpayers on a 100m $\times$ 100m grid. The map displays also three types of unsettled areas: lakes and rivers in blue, forests and parks in green and other unsettled ares in grey. Such a map gives no real impression where the Berlin taxpayers live. If we magnify the presentation scale (Figure 1b) the noise effect becomes even more pronounced. The noisy feature is somewhat dampened if we increase the gridsize to 800m in Figure 2a. The scale is the same as in Figure 1a, the number of taxpayers on a 100m $\times$ 100m grid. However, the value is equal for all 64 100m $\times$ 100m grids, which constitute one 800m $\times$ 800m grid. Note the smoothing effect on the resulting density values which reduces to the density range to 0 – 15 while in Figure 1a it is 0 – 50. Thus, the density of the choropleth is smoothed by the use of larger grid sizes. If we switch, however, to a kernel density estimate[3] the resulting map in Figure 2b clearly represents the cluster structure of the taxpayer population. Here we multiplied the density value by the area of the 100m $\times$ 100m grid to obtain expected case numbers for the grid. Therefore the scale is comparable with the scale in Figures 1a and 2a.

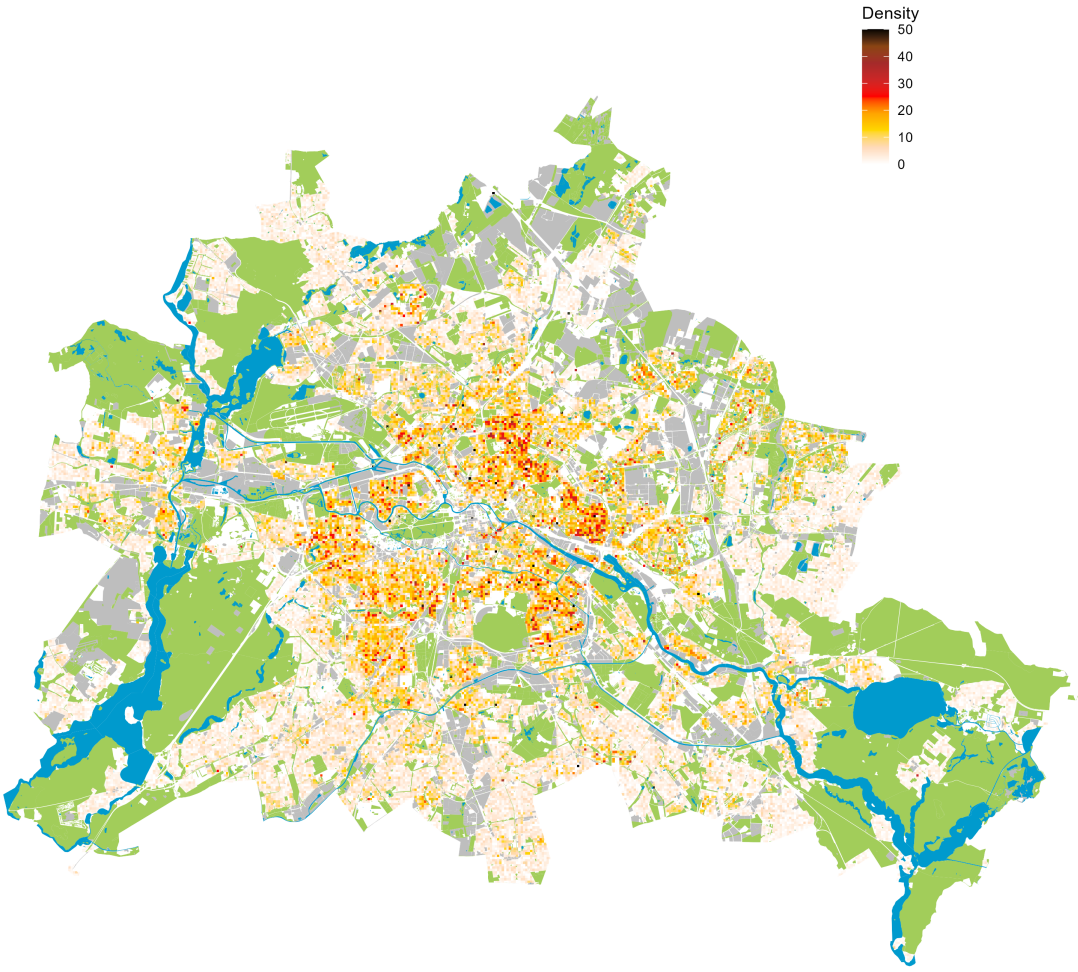
We may normalize the grid counts to a density function $f_{AGG}(x)$ and compare it with the kernel density estimate $f_{KDE}(x)$. The root mean integrated squared error (RMISE) compares the two densities across all grid points $x_\}$ $(\} = 1, \ldots, G)$ which were used for the display of the map:

$$RMISE = \sqrt{\sum_{\}} (f_{AGG}(x_\}) - f_{KDE}(x_\}))^2}$$

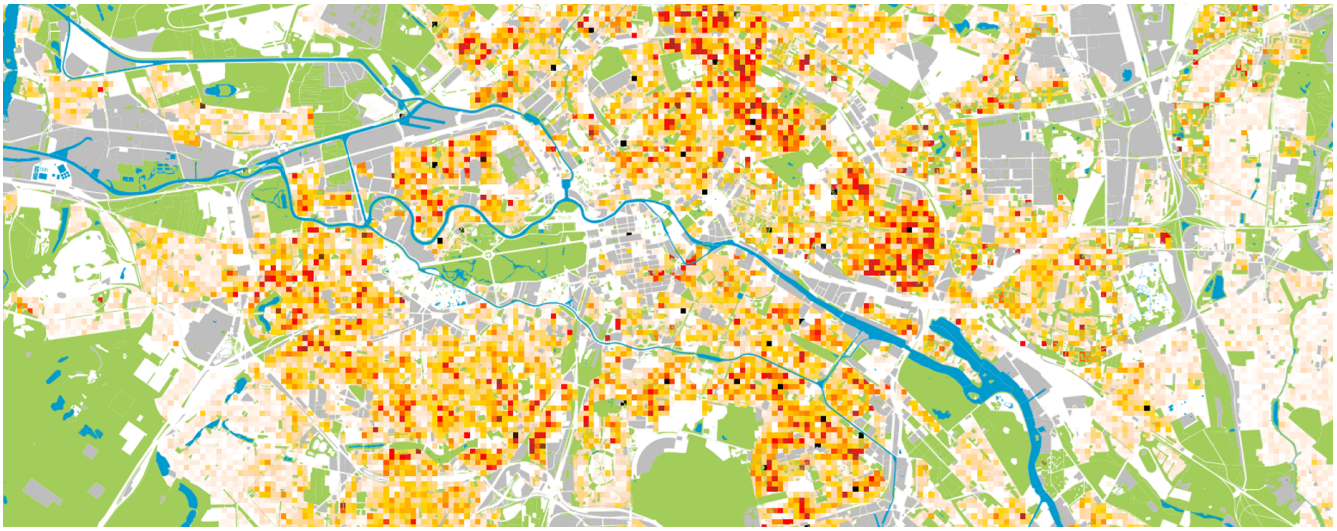
The RMISE distance of $f_{AGG}(x)$ and $f_{KDE}(x)$ is displayed for various grid lengths in Figure 3. We computed the RMISE distance for 100 replications of the sub-sampling. These replicated values created the boxes which are displayed in Figure 3. There is almost no variation of the

[3]We evaluated the density values on a 100m $\times$ 100m grid.

Figure 1: Maps of the Berlin taxpayer population (1/2).

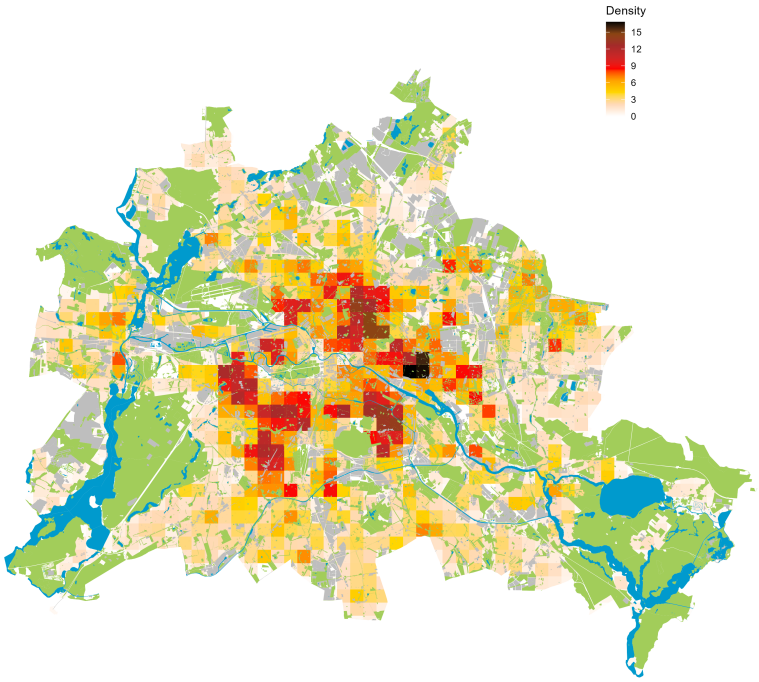


(a) Choropleth on a 100m × 100m grid.

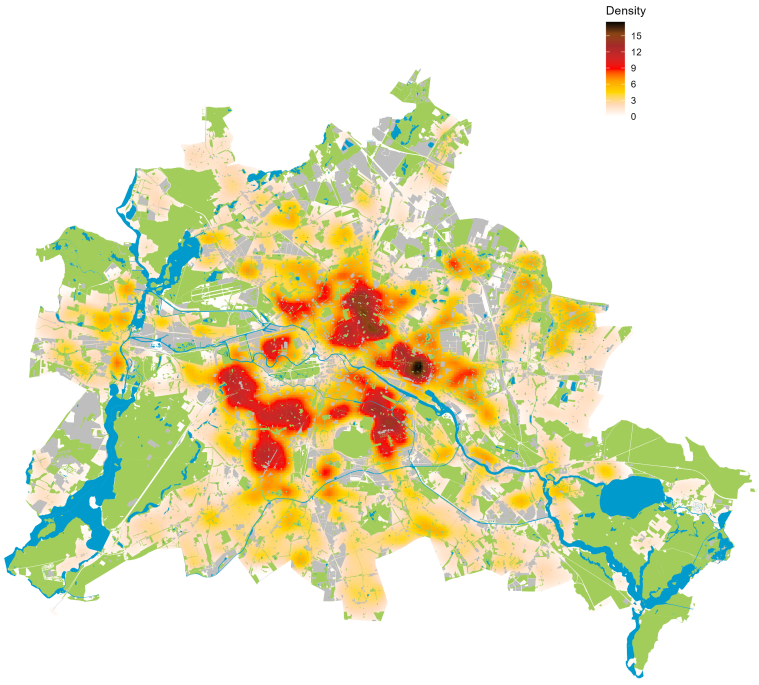


(b) Detail in enlarged scale.

Figure 2: Maps of the Berlin taxpayer population (2/2).



(a) Choropleth on a 800m × 800m grid.



(b) Kernel density estimate on the basis of the true addresses.

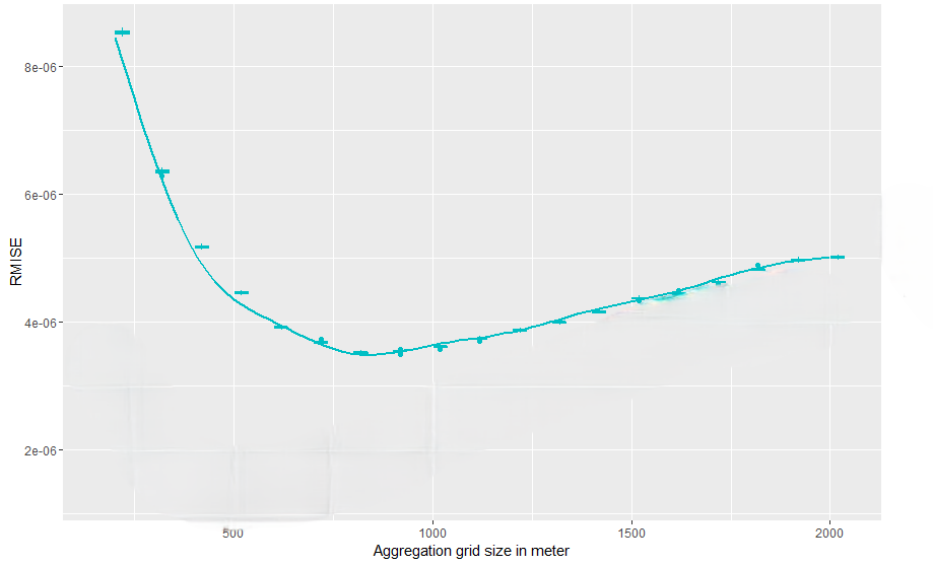


Figure 3: Impact of the grid size on the RMISE criterion

RMISE value over the replications. The line which connects the centers of the boxes is generated by a spline interpolation of the medians. It indicates that a grid of 800m $\times$ 800m is the best representation with respect to the RMISE criterion. The corresponding density map is displayed in Figure 2a.

3 Kernel density smoothing

The basic idea of the kernel density approach is the smearing of the location of an observation by a kernel function $k(x)$. The kernel function can be interpreted as a density function on $\mathbb{R}^2$ which is symmetric, i.e. $k(x) = k(-x)$, positive and normed

$$\int_{\mathbb{R}^2} k(x)dx = 1$$

The kernel density estimator is defined by:

$$\hat{f}_H(x) = \frac{1}{n \det(H)^{1/2}} \sum_{i=1}^{n} k(H^{-1/2}(x - x_i)) \tag{1}$$

where the $x_i \quad (i = 1, \ldots, n)$ are the locations of n observations and H is a positive-definite smoothing matrix which scales the distance of the i-th observation x_i and the place x where the

density is evaluated. The factor $1/(n \det(H)^{1/2})$ normalizes $\hat{f}_H(x)$ to a density over $\mathbb{R}^2$. If H is a diagonal matrix with equal entries h equation 1 reduces to

$$\hat{f}_h(x) = \frac{1}{nh^2} \sum_{i=1}^{n} k\left(\frac{(x - x_i)}{h}\right) \tag{2}$$

For notational convenience, we use the simplified scaling below.

The kernel density estimator may be interpreted as a nonparametric estimate of an underlying density f where the n observations come from, i.e. the x_i are iid observations from a distribution with density f, see, for example, Silverman (1986). There exist several kernel functions which are in use, for example, the bivariate normal distribution, the uniform distribution over an rectangle or the Epanechnikov kernel which is a radial symmetric parabola, see Silverman (1986) for details. The essential parameter, however, is the smoothing factor h. Here Wand and Jones (1994) derived a plug-in solution which minimizes an approximation of the Mean Integrated Squared Error (MISE):

$$MISE = E\left[\int_{\mathbb{R}^2} (\hat{f}_h(x) - f(x))^2 dx\right]$$

Density estimators for different populations may be written in a unique fashion by the use of indicator functions

$$\mathbb{1}_P(i) = \begin{cases} 1, & \text{if unit i belongs to population P;} \\ 0, & \text{else.} \end{cases}$$

Then the density of population P is estimated by

$$\hat{f}_P(x) = \frac{1}{h^2} \frac{1}{\sum_{i=1}^{n} \mathbb{1}_P(i)} \sum_{i=1}^{n} \mathbb{1}_P(i) k(\frac{x - x_i}{h}) \tag{3}$$

where $\sum_{i=1}^{n} \mathbb{1}_P(i) = n_P$ is the size of population P.

With this notation we can easily display population ratios. If Q is a subpopulation of population P the local ratio $r_{Q|P}(x)$ of the subpopulation Q at location x may be estimated by:

$$\hat{r}_{Q|P}(x) = \frac{n_Q \hat{f}_Q(x)}{n_P \hat{f}_P(x)} \quad (4)$$

$$= \frac{\sum_{i=1}^{n} \mathbb{1}_Q(i) k(\frac{x-x_i}{h})}{\sum_{i=1}^{n} \mathbb{1}_P(i) k(\frac{x-x_i}{h})} \quad (5)$$

if the denominator $\sum_{i=1}^{n} \mathbb{1}_P(i) k(\frac{x-x_i}{h}) > 0$, else the ratio of Q and P is set to zero.

As the population Q in the nominator is more rare than the population P in the denominator the optimal smoothing parameter for population Q is larger than for population P. Thus, in order to stabilize the ratio the larger smoothing parameter for population Q should be used.

The ratio is a special case of a mean when the dependent variable is a 0/1- variable like the membership indicator of population Q in the above case. For a general continuous variable with observation g_i at coordinate x_i we may estimate the mean of the g-values in population P at coordinate x by:

$$\hat{m}_P(x) = \frac{\sum_{i=1}^{n} g_i \mathbb{1}_P(i) k(\frac{x-x_i}{h})}{\sum_{i=1}^{n} \mathbb{1}_P(i) k(\frac{x-x_i}{h})} \quad (6)$$

This is the Nadaraya-Watson estimator (Härdle 1990, ch. 5) of the local mean $m_P(x)$ of g-values at coordinate x.

4 Statistical disclosure criteria for maps

4.1 Criteria derived from tabular data

The statistical disclosure criteria for regional maps were derived from criteria for the publication of tabular data (Hundepool et al. 2012). In this context a map is nothing else but a table with geographically arranged cells where local counts, ratios and means are displayed. A table cell which fullfills the disclosure criteria below is called a "safe area" (deJonge and deWolf 2016). These disclosure criteria were developed for choropleth maps with fixed reference areas.

There are basically three rules: The minimum frequency rule, the diversity rule and the dominance rule. The **minimum frequency rule** states that each displayed cell $j \quad (j = 1, \ldots, J)$

should have a cell count $N_j \geq \mathbf{f}$ where $\mathbf{f}$ is a predefined limit value. An exception are empty cells with $N_j = 0$. The minimum frequency rule is thought to reduce the risk to identify a person in the data set. Once a person is identified all information in the data set can be attributed to the identified person. However, such a "gain" from de-anomymization is only possible if the attacker has access to this information which is **inside** the statistical agency. But here we consider only the case that the map is published outside the agency with no access to the other variables.

The **diversity rule** refers to ratios. If a ratio r_j in cell j is near 1 or 0 then one can conclude that almost all elements in that cell belong to the population Q ($r_j \approx 1$) or do not belong to Q ($r_j \approx 0$).

The **dominance criterion** refers to the total G_j of a variable g in area j. This total should not be dominated by a contribution from a single value person. Here each single contribution $g_{i,j} \quad (i = 1, \ldots, N_j)$ in cell j should not exceed a fixed percentage $\mathbf{k}$ of the total G_j. A similar argument holds for regional percentages.

4.2 Application to maps

Since kernel maps do not know a fixed system of reference areas it is by no means clear how to apply the above criteria to a kernel density map or a Nadaraya Watson estimator. Han et al. (2019) introduced the concept of the **resolution of a map**. This is given by the grid of points $x_g \quad (g = 1, \ldots, G)$ where the kernel function is evaluated. It is the finest possible resolution for the use of a map. If a zoom-in facility is used the resolution value is a lower bound for the zoom-in. The selection of the resolution is by no means neutral with respect to the anonymity criteria. The smaller the grid size the harder it is to meet the minimum frequency criterion. The same holds for the dominance criterion: the fewer values contribute to the cell total the more pronounced is the effect of the maximum value on the total. Also cells with ratios near 0 and 1 become more frequent with declining grid size. Han et al.(2019) demonstrated the substantial effect of the resolution on the percentage of unsafe areas.

Besides the resolution one has to select also critical values which determine safe areas. Ideally, the selction of these values should be linked to the probability of the disclosure of a person.

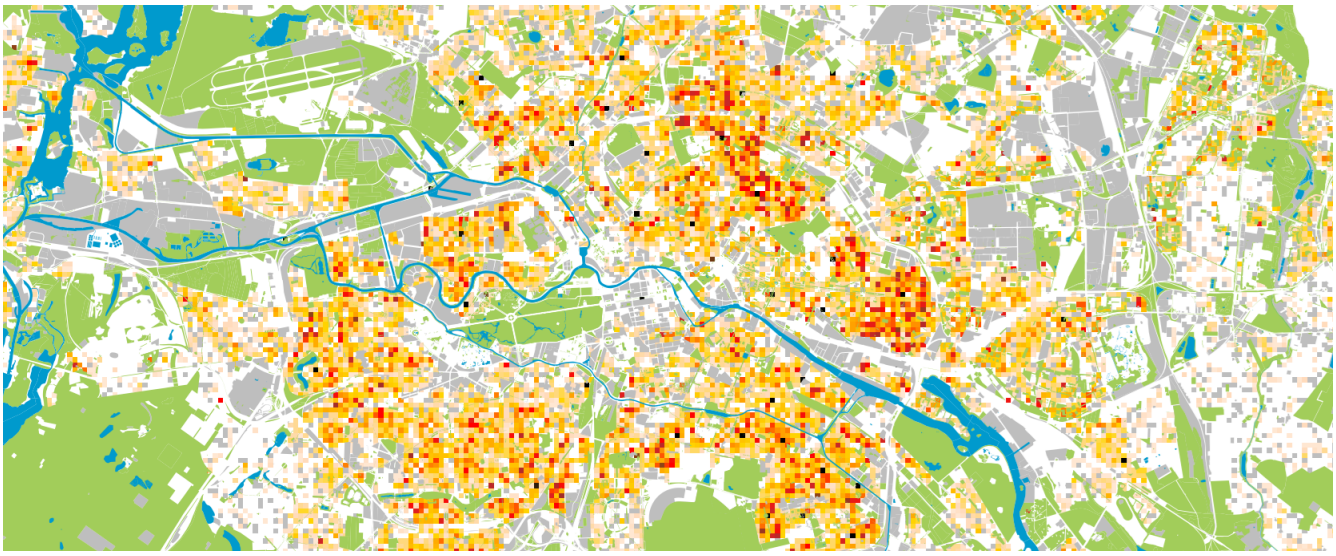
However, the calculation of disclosure probabilities is a difficult task where strong support from statistical modelling is needed, see, for example, Reiter (2005). Zhou et al. (2010) estimate disclosure risks where the original addresses are masked by spatial smoothing. They assume that the entire population enters the computation of the smoothed values. Such a condition is not met if the database comes from a survey or if only a subsample of the population is used as in the previous taxpayer example.

But even if one would be able to establish an explicit relationship of the critical values and the probability of detection one would have to answer the question: is a detection probability of, say, 10 percent tolerable? Or should it be not more than 5 percent? The answer may depend on the sensitivity of the variable of interest. In the taxpayer example it was the indicator "the person is a taxpayer", which is not really sensitive. A more sensitive indicator might be whether a person has a taxable income above the 90 percent quantile of all taxpayers. Because of the vague knowledge of the true detection risks one may select the frequency parameter **f** or the dominance value **k** by convenient choices as it is frequently done in official statistics. For example, one may chose the parameters such that the resulting percentage of unsafe areas should be moderate.

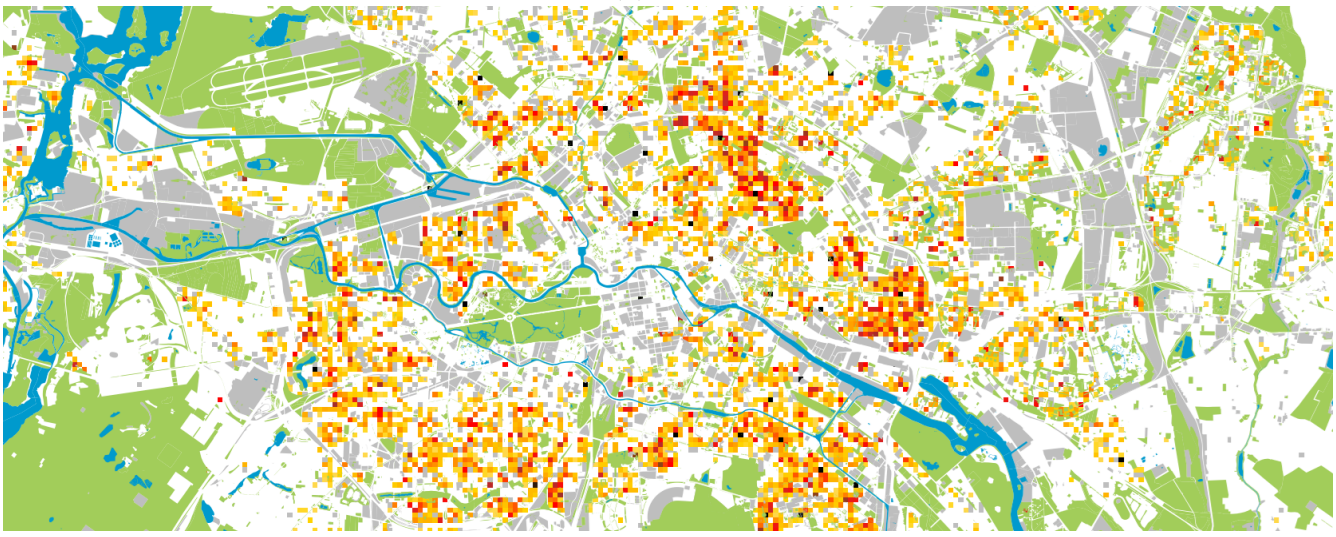
In order to demonstrate the impact of the frequency limit **f** we display the unsafe areas in the taxpayer example for a $100m \times 100m$ resolution. Again only a 10 percent sample of the original population is used, which still corresponds to 190 thousand observations. Figure 4a shows the grid cells with less than $\mathbf{f} = 3$ observations by a white coding color[4]. We have increased the scale of the map to make it easier to identify the cells which fail the minimum frequency criterion. Even in the center of the city one finds a lot of white grid cells. This analysis is repeated for $\mathbf{f} = 10$. As can be seen from Figure 4b the percentage of unsafe areas has now become so frequent, that an irregular pattern of areas remains white and it therefore hard to get an impression of the distribution of the taxpayer population.

[4]The coding of the other colors corresponds to the original map in Figure 1a.

Figure 4: Areas which fail the frequency criterion for different minimum values of **f**. Areas with case numbers below **f** are coded by white color. Grid size: $100m \times 100m$. Total number of observation: 190 thousand taxpapayers.



(a) $\mathbf{f} = 3$



(b) $\mathbf{f} = 10$

5 The smoothing effect of kernel estimates

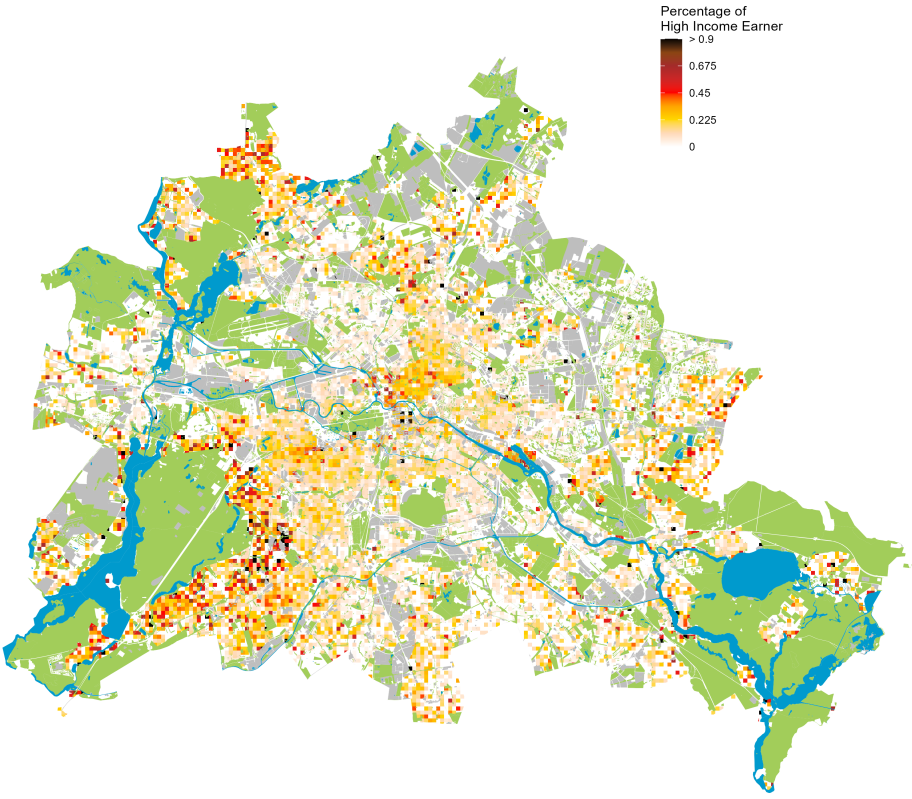
In this section we will demonstrate the smoothing effect of the kernel estimators. For this purpose we use again the 10 percent subsample of the taxpayer population. The upper 10 percent quantile of the annual taxable income of the Berlin population is around 55000 € in 2019. The population proportion of these high income persons is by definition 10 percent. We are interested in the local proportions of these high income persons.

Figure 5a displays the local proportions on a $100m \times 100m$ grid without kernel smoothing. The range of the local proportions is quite large. However, extreme values are hidden by the coding of the legend, for example, all areas with ratios higher than 90 percent are collapsed to the category $\geq 90\%$. Thus the diversity rule is preserved by the coding of the legend. The region with the most high income person ratios near the Grunewald in the south-west of Berlin is displayed in greater detail in Figure 5b. In this scale one recognizes the noisy structure of the local ratios. One also recognizes that unsettled areas are put as an additional map layer upon the grid layer. Because of the noisy structure it is hardly possible to identify regional clusters of relevant size of high income persons. Finally, Figure 6 demonstrates the smoothing effect of the kernel estimate of the ratio. This results in a substantial reduction of the range of estimated local ratios which are now not larger than 40 percent. This can be interpreted as the anonymization effect of the kernel smoothing approach. We can also see clear local clusters of areas of high income persons.

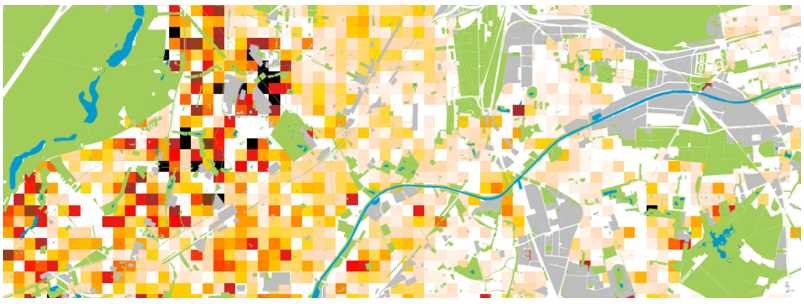
6 Combining frequency restrictions and kernel estimates in a map

One can use a publication strategy which respects on the one hand the minimum frequency rules and uses on the other hand the smoothing effect of the kernel estimates and its display of local clusters. In cooperation with the colleagues of the statistical agency who hosted the data base we were advised to suppress areas which fail the frequency rules. For the other areas we were allowed to display the kernel density estimates which were computed on the basis of **all** observations, including the observations of the unsafe areas.

Figure 5: Local proportion of high income taxpayers. 10 percent subsample of population.

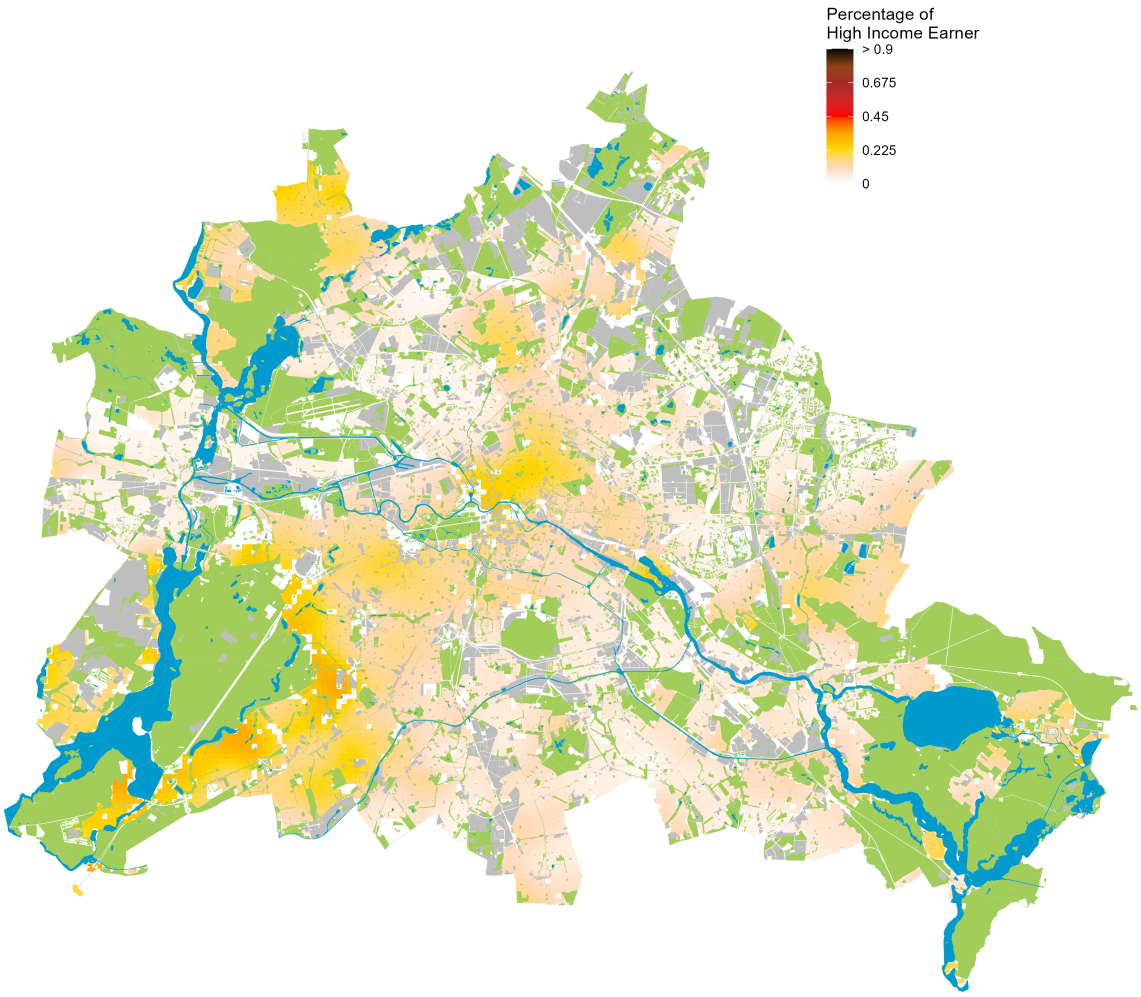


(a) Choropleth map on a $100m \times 100m$ grid



(b) Detail of choropleth map in the Grunewald-Area

Figure 6: Local proportion of high income taxpayers. 10 percent subsample of population. Kernel estimate of local ratio



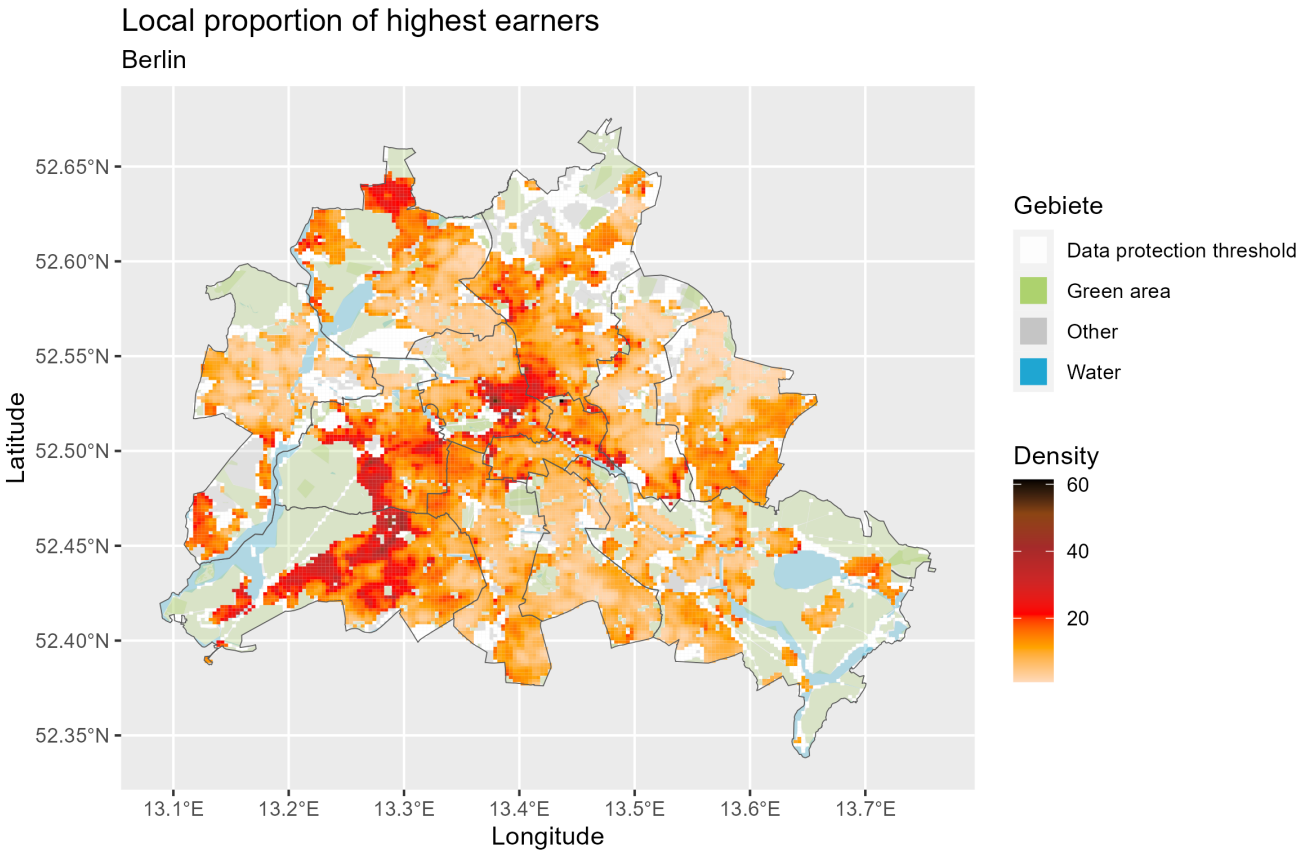


Figure 7: Local proportion of high income taxpayers in Berlin. Kernel estimation with suppressed areas in white color. Minimum frequency of taxpayers in grid cells $\mathbf{f} = 30$. Grid cell: $160m \times 250m$.

The critical parameter is the size of the area. There should be only small areas which are suppressed. This votes for small areas. On the other hand small areas are prone to have low case numbers. This votes for larger areas. We were advised to use a minimum frequency of $\mathbf{f} = 30$ taxpayers. This corresponds to a minimum frequency of 3 in case of a 10 percent subsample. For a $100m \times 100m$ this minimum frequency limit was too high to result in a reasonable number of safe areas. Therefore we did choose a larger gridsize. In Figure 7 we used a $160m \times 250m$ grid, which resulted in a moderate number of unsafe areas. Despite the map being somewhat disturbed by the suppressed areas on can still recognize clear regional clusters which are generated by the smoothing effect of the kernel estimate.

7 Other aspects of anonymization

7.1 The format of map output

The format in which a map is published has important consequences for the disclosure status of the map.

For example, the German Census-Atlas 2022[5] does not only provide interactive maps which can be zoomed-in from 10km down to 100m grids, it also provides the data files with cell frequencies and cell means at the resolution levels 10km, 1km and 100m. This already fixes the resolution level to $100m \times 100m$ grids. However, these values are not the population values, but they were subjected to output anonymization by the cell key method (Enderle and Kleber 2024). Thus there is no need for cell suppression as all modified cells are conform with the frequency limit. However, no kernel estimates are supported. One can only calculate density estimates via the grid totals by application of the Kernel Heaping algorithm which is based on the use of area totals, see Groß et al. (2017), Erfurth et al. (2022). The resulting output Kernel estimates need no further anonymization as their basis are already anonymized cell data. Note, that the computation of the kernel densities is still promising as the resulting maps display regional clusters which may not be seen from the choropleth maps.

The EUROSTAT Statistical Atlas[6] presents a diversity of maps at a 1 km grid level which can be downloaded as PDF-files. Here the corresponding data files of the cell frequencies are not offered and also the legend for the display of proportions is fixed to six intervals, thus protecting areas with extreme proportions. However, also here there are no density maps offered.

Occasionally, density plots can be found in statistical journals, see, for example, Groß et al. (2017) and Erfurth et al. (2022). Generally these maps can be downloaded. However, as a rule the authors do not publish the resolution of the density map. Also, in most instances the labelling of the density values is done by intervals. Thus it is not possible to get access to the exact density value at the grid points. As a consequence it is not possible to switch from the density value $\hat{f}_h(x_j)$ at the center of a "cell" j to an estimated number of observations $\hat{n}_j$ in the cell by using

[5] See `https://www.zensus2022.de/DE/Aktuelles/Hinweis_Zensusatlas.html`.
[6] See `https://ec.europa.eu/statistical-atlas/viewer/`.

the relationship $\hat{n}_j = N_P \hat{f}_h(x_j)\Delta$. Here Δ is the area of the grid cell and N_P is the total of the population of interest. Thus any reference to the minimum frequency rule is not possible for users of the map.

Besides, the inference to the true case numbers in the population is subject to statistical variance. The asymptotic variance of a kernel density estimate with a bivariate normal kernel function[7] is:

$$Var(\hat{f}_h(x)) = \frac{1}{4\pi h^2 n} f(x) \tag{7}$$

where n is the sample size of the kernel estimate. Thus, the asymptotic variance of $\hat{n}_j$ can be estimated by:

$$\begin{aligned} \hat{Var}(\hat{n}_j) &= \frac{N_P^2 \Delta^2}{4\pi h^2 n} \hat{f}_h(x_j) &(8)\\ &= \frac{N_P \Delta}{4\pi h^2 n} \hat{n}(x_j) &(9) \end{aligned}$$

Hence the standard deviation of $\hat{n}(x_j)$ is proportional to its root value. If all population values are used for the kernel estimate, i.e. $N_P = n$, and if the area between the grid point Δ is equal to h^2 the proportionality factor simplifies to $1/(4\pi)$. Thus for $\hat{n}_j = 3$ we obtain a standard deviation of 0.488. In this case the estimated case number would range between 2 and 4 if we take two times the standard error as a confidence region of the true cell frequency. Thus we are not sure whether a minimum frequency rule with $\mathbf{f} = 3$ applies for the cell or not. In case of a lower sample size n the confidence interval may even become wider.

From this point of view, one could argue that a plain kernel smoothing map as a PDF file without reference to the resolution and without the exact kernel density values is a safe map. However, the good advice from inside the statistical agency to suppress cells with low frequency creates some extra safety which does not destroy the benefits from detecting regional clusters by the kernel smoothing approach.

[7] See `https://en.wikipedia.org/wiki/Multivariate_kernel_density_estimation`.

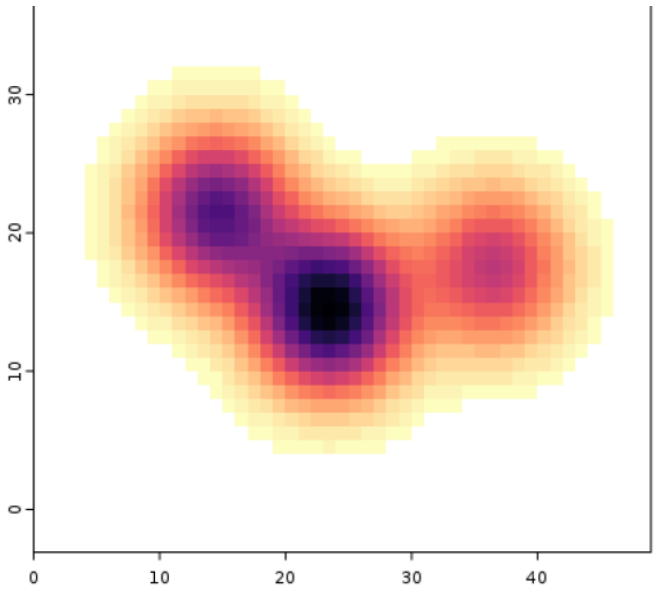


Figure 8: A group of three isolated points with overlapping kernel functions

7.2 Specific anonymization issues of kernel smoothing

As shown above, the concept of anonymization based on rules for observed frequencies does not really fit to kernel smoothing. However, there are some anonymisation issues which are specific for kernel smoothing estimates.

7.2.1 Isolated observations

The kernel approach has some difficulty to anonymize isolated points or even a small group of isolated points. Figure 8 shows a situation where 3 kernel functions partially overlap. If the sample size of our data base is large, which implies small values of h and there are areas with low population size, the above situation may occur. Because of the symmetric nature of the kernel function it is very easy to reconstruct in Figure 8 the true location of the three observations. This would violate the anonymity. In real situations the degree of overlap depends on the smoothing parameter h. Note, that in our empirical examples such a case did not occur.

For this reason one has to build-in some protection against the uncovering of geographically isolated observations. Here one might look for kernel procedures with adaptive smoothing parameters, see Sadiq et al. (2022). The adaptive component would increase the smoothing factor in regions with few observations. However, this approach has to be still developed in more detail.

7.2.2 De-anonymization of individual values

In cases of populations of moderate size, like firms, the exact location of the units may be known from open sources. However, the output variable g_i of unit i is unknown and of interest. The Nadaraya Watson estimator computes a locally weighted average of the original g-values by

$$\hat{g}_i = \sum_j \frac{k(\frac{x_i - x_j}{h})}{\sum_l k(\frac{x_i - x_l}{h})} g_j \tag{10}$$

This is a linear equation system where the vectors $g = (g_1, \ldots, g_N)'$ and $\hat{g} = (\hat{g}_1, \ldots, \hat{g}_N)'$ are connected by multiplication of a matrix $K = (k_{ij})_N^N$ with

$$k_{ij} = \frac{k(\frac{x_i - x_j}{h})}{\sum_l k(\frac{x_i - x_l}{h})}$$

Hut (2020) has proven that the matrix K is invertible for a normal distribution kernel function. Thus, with knowledge of the $x_i \quad (i = 1, \ldots, N)$ and the kernel function k and the used smoothing value h one may simply compute the inverse of K, multiply it with $\hat{g}$ and obtain the unknown values of g.

From a practical point of view there are frequent situations where the above strategy will not work. In the taxpayer example one would have to invert a matrix of dimension $N = 1.9$ million, which is simply an enormous calculation task. As many taxpayers have the same address all rows of taxpayers with the same address are identical in the matrix K. This leads to a rank defect and prevents the inversion of K. If we have a register of all Berlin addresses at disposal then not every address belongs to a taxpayer. This may result in more rows than columns in the K matrix leading to an over-identified linear equation system.

For cases where K is invertible Hut (2020) proposes to put some stochastic noise on the kernel estimates $\hat{g}(x_i)$ such that

$$max_{i=1,\ldots,N} \left\{ Prob \left(\left| \frac{\hat{g}(x_i) - g(x_i)}{g(x_i)} \right| < \frac{p}{100} \right) \right\} \leq \alpha \tag{11}$$

For a normal noise function and invertible matrix K Hut (2020) presents an explicit relationship of the parameters p and α and the obtained anonymization probability in Equation 11.

7.2.3 The selection of the smoothing parameter h

If we concentrate on a fixed kernel function the smoothing parameter h is the only parameter of the kernel estimates. Here, the use of the bivariate normal density is frequent.

It is instructive to see that the selection of h is also sensitive with respect to the anonymization effect. For $h \to \infty$ the true locations are smeared over the entire region of interest. In the case of density estimation this results in a uniform density over the entire region. In the case of the estimation of means the Nadaraya Watson estimator is simply the average over the entire area. If proportions are of interest the limit is given by the overall ratio over the area.

In the other extreme, i.e. $h \to 0$, the estimated density is given by the set of Dirac functions at the exact locations of the observations. In case of a normal kernel function Hut (2020) has proven that the Nadaraya Watson estimator $\hat{g}(x)$ at location x converges to the value of $g(x_i)$ of the observation which is nearest to x. As a conserquence the ratio estimate $\hat{r}(x)$ converges to 0 or 1, depending on the population status of the nearest observation x_i. If observation i belongs to population Q then $\hat{r}(x) = 1$ results and $\hat{r}(x) = 0$ else. Thus, for $h \to 0$ the kernel smoothing reveals the exact positions, the population status and the individual values of the observed units. Consequently, the data protection is zero. For the opposite case $h \to \infty$ the statistical use of the geo-coordinates is zero, as we do not not use this geo-information.

The standard reasoning of anonymization uses a bipolar scale ranging from low anonymization with high data utility to a high anonymization with low data utility, see, for example, de Jonge and de Wolf (2016). Usually it is argued, that a maximum risk limit should be strictly respected. Below the limit the utility of the data should then be maximized. Cox et al. (2011) have argued against this model as to simplistic for applications. As a rule, the risk and the utility are hard to specify in concrete statistical terms. In the case of kernel smoothing there is a direct relationship of the anonymization parameter h and the use of $\hat{f}_h$ as an estimate of the true population density f. A standard selection criterion for h is the minimization of the MISE criterion, see Wand and Jones (1994). Thus we are in a situation were the selection of h optimizes as well the utility of the obtained map at a price of a moderate anonymization effect.

8 Concluding remarks

Due to the smearing effect of the kernel density approach the impact of single values on the resulting map is largely reduced. Thus the output map may be regarded as an anonymized output.

In general it is difficult to apply anonymization criteria which were formulated for tabular analyses to the output of kernel smoothing routines. If the population is large as in the taxpayer example a simple subsample, say 10 percent, of the original population may be enough to safeguard against de-anonymization. Alternatively one may suppress the display of regions where the case numbers at the selected grid resolution fail to meet the chosen minimum frequency criterion. However, the computation of the kernel estimate has still to include the information of the observations in the suppressed areas. In our empirical example with partly suppressed areas the resulting map remains still informative with respect to the displayed local clusters of high income taxpayers. However, the higher the resolution of the map is, the larger becomes the fraction of suppressed areas. So the resolution of the displayed map is essential here.

More attention should be paid to the output format of the map. If the map is supported by machine readable files with the computed density values then, at least theoretically, it is possible to de-anonymize the original values of the observations. However, in our empirical example with high income taxpayers, such a de-anonymization was not possible. Also the labeling of the displayed means by intervals is an efficient method to hide dominance effects of single observations. Thus the output format of the map may turn out to be a more efficient anonymization tool than the standard application of frequency and dominance rules for tabulations.

We did not consider here output anonymization, for example, the cell key method, as it is designed for the display of tables like choropleth maps. As we did show in our examples, choropleth maps perform poorly to display regional clusters. This feature becomes even worse with smaller grid sizes. Unless the cell frequencies refer to administrative areas of interest, the kernel smoothing maps should be preferred to choropleth maps.

There are even further advantages from the computation of the density map. The computed densities offers the possibility to display high density areas which are independent from any area system. Such high density areas may be used to fix regional clusters which characterize the

concentrations of certain sub-populations, for example, students (Groß et al. 2020) or voters of a political party (Erfurth et al. 2022).

9 Acknowledgement

This work was part of the EU-funded Competence Cluster Anonymization of Integrated and Geo-referenced Data (AnigeD) (https://www.destatis.de/DE/Ueber-uns/AnigeD/_inhalt.html). Lorena Gril was financed under Research Contract 16KISA097

References

Armstrong,M; Rushton,G; Zimmerman,D (1999): Geographically masking health data to preserve confidentiality, Stat. Med. 18, 497- 525

Cox, L; Karr, A; Kinney, S (2011): Risk-Utility Paradigms for Statistical Disclosure Limitation: How to Think, But Not How to Act. International Statistical Review, 79, 160 - 183.

de Jonge, E. (2022): The *sdcSpatial* Package

https://cran.r-project.org/web/packages/sdcSpatial/index.html

de Jonge, E; de Wolf, P.-P. (2016): Spatial Smoothing and Statistical Disclosure Control. In: Domingo-Ferrer and Pejic-Bach (Eds), PSD Conference 2016, pp. 107- 117. https://doi.org/10.1007/978-3-319-45381-1_9

Destatis (2024): Der Zensus Atlas 2022. https://atlas.zensus2022.de

de Wolf, P.-P.; de Jonge, E. (2018): Safely Plotting Continuous Variables on a Map. In: Domingo-Ferrer and Montes (Eds), PSD Conference 2018, pp. 347 - 359. https://doi.org/10.1007/978-3-319-99771-1_23

Enderle, Tobias und Kleber, Birgit (2024): Geheimhaltung mit der Cell Key Methode im Zensus 2022. Statistisches Bundesamt WISTA Nr. 6-2024.

https://www.destatis.de/DE/Methoden/WISTA-Wirtschaft-und-Statistik/_inhalt.html

Erfurth, K; Groß, M; Rendtel, U; Schmid, T. (2022): Kernel density smoothing of composite spatial data on administrative area level: A case study of voting data in Berlin. AStA Wirtschafts- und Sozialstatistisches Archiv, 16, 25 - 50,

https://doi.org/10.1007/s11943-021-00298-9.

Gril,L; Steinkemper, L.; Groß, M; Rendtel,U (2024): The use of the R-package *kernelheaping*. The R journal, To appear.

Groß, M.; Kreutzmann, A.-K.; Rendtel, U; Schmid, T.; Tzavidis, N. (2020): Switching between different area systems via simulated geo-coordinates: A case study for student residents s in Berlin. Journal of Official Statistics, 36, 297 – 314, http://dx.doi.org/10.2478/JOS-2020-0016

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.

https://academic.oup.com/jrsssa/article/180/1/161/7068203

Härdle, W. (1991): Smoothing techniques. Springer Series in Statistics, Springer New York.

Haddam, S; Schmid, T; Simm, J (2020): Kleinräumige Prädiktion von Bevölkerungszahlen basierend auf Mobilfunkdaten aus Deutschland. In: Klumpe et al. (Hrsg.), Qualität bei zusammengeführten Daten, Schriftenreihe der ASI – Arbeitsgemeinschaft Sozialwissenschaftlicher Institute,

https://doi.org/10.1007/978-3-658-31009-7_3

Han, S; de Wolf, P.-P.; de Jonge, E (2019): Comparing methods of safely plotting variables on a map. PSD Conference 2019,

Hossain, J (2023): Statistical Estimation and Inference with Aggregated and Displaced Georeferenced Data. PhD Thesis University Southampthon

https://eprints.soton.ac.uk/484015/

Hundepool,A; Domingo-Ferrer, J; Franconi, L; Giessing, S; Schulte Nordholt, E; Spicer, E; de-Wolf, P.-P (2012): Statistical Disclosure Control, Wiley.

https://doi.org/10.1002/9781118348239

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Naylor, J., Schulte Nordholt, E., Seri, G., de Wolf, P.-P., Tent, R., Mlodak, A., Gussenbauer, J., and Wilak, K. (2024). Handbook on statistical disclosure control, second edition. STACE project WP2 Deliverable D2.10. https://sdctools.github.io/HandbookSDC.

Hut, D. (2020): Statistical Disclosure Control When Publishing on Thematic Maps Master Thesis University Twente,

https://essay.utwente.nl/82311/

Risk-Utility Paradigms for Statistical Disclosure Limitation: How to Think, But Not How to Act Lawrence H. Cox, Alan F. Karr, Satkartar K. Kinney

Lagonigro,R; Oller, R; Martori, J (2021): AQuadtree: an R Package for Quadtree Anonymization of Point Data, The R Journal 12(2) https://doi.org/10.32614/RJ-2021-013

Meindl, B 2023: introduction to the *cellKey* package. https://cran.r-project.org/web/packages/cellKey/vignettes/introduction.html

Eurostat (2025): Guidelines for Statistical Disclosure Control Methods Applied on Geo-Referenced Data. https://ec.europa.eu/eurostat/en/web/products-manuals-and-guidelines/w/ks-01-25-005

Reiter, J (2005): Estimating risks of identification disclosure in Microdata. J. Amer. Statist. Assoc. 100, 1103 -1112 https://doi.org/10.1198/016214505000000619.

Rendtel,U; Schmid, T (2024): On the estimation of smooth maps from regional aggregates via measurement error models: A review. In: Knoth, Okhrin, Otto (Eds.): Advanced Statistical Methods in Process Monitoring, Finance, and Environmental Science. Essays in Honour of Wolfgang Schmid. Springer ISBN 978-3-031-69110-2 https://link.springer.com/book/9783031691102

Sadiq, Dent and Wysocki (2022): Flexible and Fast Estimation of Binary Merger Population Distributions with Adaptive KDE. https://arxiv.org/pdf/2112.12659

Silverman, B (1986) : Density Estimation for Statistics and Data Analysis, Routledge, https://doi.org/10.1201/9781315140919

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

Wikipedia "Multivariate Kernel density estimation": https://en.wikipedia.org/wiki/Multivariate_kernel_density_estimation

Wang,Z; Liu, l; Zhou,H; Lan,M (2019) : How is the confidentiality of crime locations affected by parameters in kernel density estimation? (ISPRS International Journal of Geo-Information,8(12):544, https://doi.org/10.3390/ijgi8120544

Zhou, Y; Dominici, F; Louis, Th (2010): A Smoothing approach for Masking Spacial Data, Annals of Applied Statistics, 4 (3), 1451- 1475, https://doi.org/10.1214/09-AOAS325

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

2025 erschienen:

2025/1 Coleman, Winnie und Dieter Nautz: Asymmetric Inflation Target Credibility
Economics

2025/2 Suesse, Marvin und Theocharis Grigoriadis: Financing late industrialization: evidence from the State Bank for the Russian Empire
Economics

2025/3 Schurig, Tim; Sukumar Munshi, Christian Buggedei, Jürgen Eils, Victor Ziehe, Arthur Kari und Martin Gersch: Dokumentation vom Health-X Meilenstein 14: Umsetzung ausgewählter Geschäftsmodelle
Information Systems

2025/4 Deparade, Darius; Lennart Jarmolinski und Peter N.C. Mohr: Behavioral Interventions, Tax Compliance and Consequences on Inequality
Economics

2025/5 Prummer, Anja und Francesco Nava: Divisive by Design: Shaping Values in Optimal Mechanisms
Economics