Gril, Lorena; Rendtel, Ulrich

Working Paper

Mapping high-income taxpayers in Berlin using kernel-smoothed proportions from aggregated georeferenced data

Discussion Paper, No. 2026/2

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

Suggested Citation: Gril, Lorena; Rendtel, Ulrich (2026) : Mapping high-income taxpayers in Berlin using kernel-smoothed proportions from aggregated georeferenced data, Discussion Paper, No. 2026/2, Freie Universität Berlin, School of Business & Economics, Berlin, https://doi.org/10.17169/refubium-51220

This Version is available at:
https://hdl.handle.net/10419/336787

Mapping High-Income Taxpayers in Berlin Using Kernel-Smoothed Proportions from Aggregated Georeferenced Data

Lorena Gril
Ulrich Rendtel

School of Business & Economics
Discussion Paper

Economics

2026/2

Mapping High-Income Taxpayers in Berlin Using Kernel-Smoothed Proportions from Aggregated Georeferenced Data

Lorena Gril[1,2], Ulrich Rendtel[1]

Abstract

The rare access to exact official geocoordinates opens new methodological possibilities for analyzing highly sensitive tax data. We explore their visualization potential and systematically evaluate aggregation as an anonymization strategy, with particular attention to its methodological and analytical implications. For an analysis of high-income taxpayers in Berlin, Germany, the focus is on the presentation of regional shares. In addition to frequency maps, smoothed representations using kernel density estimation are analyzed in particular, and their cartographic characteristics are discussed. Due to the high sensitivity of individual-level data, such data are generally not published, which is why anonymization is required in official statistics. This applies in particular to the group of high-income taxpayers. Using exact data as a gold standard makes it possible to systematically analyze the distortions caused by aggregation, one of the most commonly used anonymization methods in official statistics. In order to correct these distortions, a measurement error model is employed that explicitly accounts for the aggregation process and produces smoothed kernel density estimates for interpretable cartographic representations. In addition, the measurement error model is linked with census information to demonstrate a realistic application scenario. Local and global error measures are intended to empirically substantiate the improvement achieved through the use of the measurement error model.

Keywords: Anonymization, measurement error model, kernel density estimation, income tax data, census information

1 Introduction

The spatial dimension of data opens completely new perspectives for official statistics: geolocation data, understood as the assignment of demographic/survey characteristics to geoinformation such as addresses, makes it possible to map social processes – from population density to mobility patterns and land use – on a small scale. Spatial analyses can thus provide a reliable evidence base for political and administrative decision-makers, cf. Li et al. (2016). Satellite-based remote sensing data, for example, have long been used to analyze spatial structures, while their use in official statistics has been increasingly taken into account in recent years. Georeferenced administrative, survey and mobile phone data enable the integration of more precise estimates of key parameters and nuanced analyses of spatial disparities, cf. Bensmann et al. (2020) and Ricciato and Coluccia (2023). Building on this potential, our application enables the analysis of the spatial distribution of high-income taxpayers in Berlin using georeferenced wage and income tax statistics, thereby providing detailed insights into regional income structures and existing disparities. At the same time, such analyses place high demands on methodology, visualization and data protection, as confirmed by the literature on geolocated statistical data, cf. Schweers et al. (2016).

Georeferenced data only provides in-depth insights into spatial structures and processes when visualized carefully and methodically. Displaying geolocations as points on a map is not recommended, as single observations may enable inferences about individuals. Furthermore, the interpretation is difficult, especially when there are many data points, see Baddeley, Rubak, and Turner (2015). Both are true for highly sensitive information, such as the wage and income tax statistics for Berlin with over two million data points, as discussed below. For a general overview

[1]Free University Berlin, Department of Economics - Chair of Applied Statistics, Garyst. 21, 14195 Berlin, Germany
[2]Amt für Statistik Berlin Brandenburg, Alt-Friedrichsfelde 60, 10315 Berlin, Germany

of the wage and income statistics, we refer to the website of Statistisches Bundesamt (Destatis) (2025).

A fundamental method of spatial representation is discretization of the area of interest into grid cells. According to article 10 of the Federal Statistics Act (*Bundesstatistikgesetz*, BStatG), grid cells measuring one hectare (100 $\times$ 100 meters (m)) may be used for the permanent regional allocation of survey characteristics and for internal statistical analyses, cf. Brenzel and Gebers (2020). An illustrative example is the 2022 Census Atlas in Germany, where census results are visualized in an interactive map using grid cells with edge lengths of 10 kilometers (km), 1 km, and 100 m, depending on the zoom level. Grid representations can lead to data protection problems, especially in cells with few people. In the 2022 Census Atlas, data confidentiality was ensured at all grid levels. Furthermore, zooming in on their website more closely can lead to interpretation difficulties, cf. Statistisches Bundesamt (Destatis) (2024).

In addition to grid-based representations, a kernel density estimation (KDE) can be calculated based on the point data, cf. Diggle (1985) and Silverman (1986). KDE smoothes the point distribution, highlights spatial concentrations and facilitates the identification of hotspots and regional trends. In the literature, KDE is recommended for the creation of easily interpretable, smoothed maps that reveal supra-regional clusters, cf. Baddeley et al. (2015). A prerequisite for the method is exact geolocation of the individual points, as inaccuracies, e.g. from aggregation, can distort the density estimation and impair spatial interpretation. Aggregation of data means the combining of geocoordinates and their associated values into larger geographical units. As already noted by Scott and Sheather (1985), the magnitude of the bias depends on the degree of data aggregation. In the bivariate case, the simulation studies by Groß et al. (2016) demonstrate this effect very clearly.

For reasons of data protection, individual-level data is generally not published in official statistics. Individuals must be protected by appropriate anonymisation methods, with the specific implementation typically being decided on a case-by-case basis. In practice, however, various procedures have been established to ensure the confidentiality of data, cf. Eurostat (2025). The aggregation of data is a proven method of making information publicly available while significantly reducing the risk of re-identification of individuals. The use of large administrative areas for aggregation is advantageous for data availability, cf. Rushton et al. (2006). As confirmed by B. Wilson, N. Wilson, and Martin (2021) and Burian, Zapletal, and Pászto (2022), among others, administrative boundaries, such as municipalities or districts, are commonly used as aggregation units. In addition, methods can be employed to construct aggregates in such a way that the population is distributed as evenly as possible across aggregation units, which is necessarily the case for administrative boundaries. In both cases, the interpretability of the underlying point map is lost. The data is presented using choropleth maps, which assign a value to the entire aggregate, leading to abrupt transitions at the aggregation boundaries. The associated problems are exacerbated with increasing size of aggregates, cf. Rushton et al. (2006).

Aggregation systematically shifts the true geolocations to the centroids, which means that the use of smoothing methods such as KDE leads to distorted estimates and creates a conflict between aggregation required for data protection and analytical significance. Once an official body has published aggregated data, the question arises as to whether, taking into account the known measurement errors caused by the aggregation process, alternative forms of representation other than choropleth maps would allow for a more accurate representation.

It is by no means clear how the original observations are spatially distributed within the units when geolocations are aggregated to area-level units, which one would need to know in order to obtain smoothed representations while avoiding distortions. However, since the aggregation process is known, it can be interpreted as a measurement error, which allows the use of advanced methods for estimating latent geocoordinates. Building on this approach, Groß et al. (2016) developed the so-called kernel heaping method, which uses a partially Bayesian framework in which the unknown true values are treated as latent parameters and the aggregation process is explicitly modeled

in order to obtain corrected kernel density estimates, cf. Groß and Rendtel (2016) and Groß et al. (2016).

Our use case offers a unique opportunity for a systematic analysis based on exact official geocoordinates, providing access to data that has so far been rare and enabling new methodological insights. Hence, questions about visualisation and the effect of aggregation are addressed using our access to approximately two million wage and income tax records from the 2019 tax cohort – including access to their exact addresses – provided by the Berlin-Brandenburg Statistics Office. The focus is on analysing the spatial distribution of high-income taxpayers (HITP), which can be represented in an interpretable manner as regional proportions. Information on income taxpayers (ITP) with particularly high incomes is particularly sensitive and worthy of protection. Therefore, access to exact geolocations for this data offers, on the one hand, the possibility of a comprehensive analysis of the visualisation potential and associated challenges. On the other hand, it opens up the possibility of a systematic analysis of aggregation and its loss of information. Furthermore, the measurement error model that takes aggregation into account, proposed by Groß et al. (2016), aims to enable smoothed representations. Having exact geocoordinates from the tax cohort serves as the gold standard in the context of measurement errors due to aggregation. Furthermore, the proposed method is extended by linking it to an auxiliary source of information in the form of the German census grid data.

The article is structured as follows: In Section 2, aggregation is defined as a method of anonymising exact geocoordinates. Based on exact geocoordinates, a kernel density estimation, described in Section 3.1, can be made. However, when data are aggregated, a smooth representation using KDE is biased. Therefore, a measurement error model that takes aggregation into account is introduced in Section 3.2. Since the focus of the application is on the proportions of HITP in Berlin, the initial measurement error model must be extended to proportions, which will be done in Section 3.3. Before an systematic analysis on the wage and income tax data is performed, first the data source is briefly explained in Section 4.1, and then the visualisation potential of the data and the motivation of proportional maps are discussed in Section 4.2. Section 4.3 applies aggregation to the data, analyses the problems involved, and uses the measurement error model for smooth estimates and representations. In Section 4.3.1, the analysis is based on areas that vary systematically in size, and in Section 4.3.2, a realistic scenario is considered in which the aggregation is based on administrative areas and the link with census data is tested.

2 Aggregation

Measures are taken to anonymize data in order to prevent the re-identification of individual data in the context of official data or to minimize the risk as much as possible. When geocoded data is available, there are two options for anonymization. On the one hand, the coordinates of individuals can be geomasked, i.e. the original coordinate X_i is assigned an error term ν_i, which follows a certain distribution. On the other hand, each geocoordinate is linked to associated attributes, such as socio-demographic characteristics (e.g., household income). It is these associated values, rather than the geocoordinate itself, that may be modified, for example through rounding or other transformations. We focus on the former, the anonymization of the location.
A widely used method of changing an individual's geolocation is aggregation. Here, several geocoordinates that fall within a certain area are combined.

The simplest way of aggregation is to round the exact geocoordinates to a rectangle, i.e., separately along each coordinate. Let $M_a = (M_{a1}, M_{a2}),\ a = 1, \ldots,\ A$ be the centers or rounding points of aggregation rectangles, with side lengths r_1 along the horizontal axis and r_2 along the vertical axis. Then the anonymized point for X_i is given by $W_i = M_{i,a} = X_i + \nu_i$ if $X_i \in (M_{a1} - 0.5 \cdot r_1, M_{a1} + 0.5 \cdot r_1) \times (M_{a2} - 0.5 \cdot r_2, M_{a2} + 0.5 \cdot r_2)$. Aggregation by rounding results in a uniform, coarse grid over the area of interest. The structure of the underlying rectangle is denoted by P_a and can also be interpreted as a polygon.

In the field of official statistics, administrative boundaries are usually used, for example, the national border as the outermost boundary and finer subdivisions into federal states, such as provinces, municipalities or post-codes. This results in more complex structures for the aggregation areas, but – similar to rounding – points to be anonymized that lie within the same area are grouped together and moved to the center of the area. Let us assume an administrative system with a total of A areas. If X_i, the point to be anonymized, lies in the polygon P_a of area a, then it is placed in the center of P_a, i.e. $W_i = M_{i,a} = X_i + \nu_i$.
It should be noted that the anonymized location depends on both the exact location and the area system. When presenting aggregated data, for example as a map or table, the type of anonymization, i.e. the underlying area system, is known. This knowledge should be used to achieve a better presentation of the aggregated data.

3 Measurement error model accounting for the anonymization process

The evaluation of geocoordinates requires structured analysis in order to make the complexity of the data manageable, whereby spatial patterns often only become visible through the use of smoothing procedures. Assuming exact geocoordinates, kernel density estimation proves to be a suitable method for the efficient identification and representation of regional patterns.

3.1 Multivariate kernel density estimation

Multivariate kernel density estimation (KDE) represent a non-parametric approach to estimating the probability distribution of continuous random variables. They generate smooth density estimation and can be understood as a generalization of classical histograms. In a spatial context, the two-dimensional case is relevant. For a sample $X = (X_1, \ldots, X_n)$ with sample size n consisting of exact geocoordinates $X_i = (X_{i1}, X_{i2})$ representing longitude and latitude, the bivariate random variables with unknown density function $f(x)$ yield the kernel density estimators

$$\hat{f}_H(x) = \frac{1}{n|H|} \sum_{i=1}^{n} K\big(H^{-1}(x - X_i)\big), \tag{1}$$

where $K(\cdot)$ denotes a multivariate kernel function, H is a symmetric, positive definite bandwidth matrix, and $|H|$ is its determinant. In the case of the Gaussian kernel, the bandwidth matrix forms a bell-shaped weighting that is centered around the observations X_i. The choice of bandwidth matrix H is crucial for the quality of the estimation. Bandwidths that are too small lead to an unstable bumpy or noisy density estimation, while those that are too large lead to an excessively smoothed density estimation. Various methods for bandwidth selection are discussed in the literature, cf. Silverman (1986) and Izenman (1991).

Due to the construction of the kernel, kernel density estimates within bounded areas may yield positive values in peripheral or uninhabited regions. Parts of the kernel density estimate lie outside the area of interest, which distorts the density in peripheral regions. C. M. Jones (1993) suggests restricting the estimation to the area defined by $\mathcal{S}$ and using scaling factors w in the kernel density estimation to control the fit within the boundaries. For a sample point X_i, the scaling factor can be approximated by its discretized equivalent

$$w_i = \int_{\mathcal{S}} \frac{1}{|H|} K\big(H^{-1}(X_i - y)\big)\, dy \tag{2}$$

$$\approx \sum_{z \in \mathcal{S}} \frac{1}{|H|} K\big(H^{-1}(X_i - z)\big) \Delta_1 \Delta_2. \tag{3}$$

First, w_i is defined analytically as an integral over the area of interest $\mathcal{S}$ and then approximated by discretizing

$\mathcal{S}$ so that z denotes an evaluation point of the discretised area of interest and Δ_1 and Δ_2 the edge lengths of the corresponding grid cell. The weighting factor can be used in a weighted kernel density estimation to prevent the estimation from smearing into uninhabited areas or areas outside the study area and the density at the edge from being underrepresented. The simple sum in equation 1 is replaced by a weighted sum, yielding the weighted estimator

$$\hat{f}_H(x) = \frac{1}{n|H|} \sum_{i=1}^{n} w_i K\big(H^{-1}(x - X_i)\big), \tag{4}$$

since not every observation contains the same amount of information about the underlying density, cf. Hall and Turlach (1999).

In this work, the plug-in method according to Wand and C. Jones (1994) is used for bandwidth determination, which is particularly notable for its computational efficiency within the measurement error model (cf. Section 3.2). This was demonstrated in the simulation study by Gril et al. (2025).

3.2 Measurement error model for aggregated geocoordinates

Equation (1) illustrates that knowledge of the exact geocoordinates $X = \{X_1,\ X_2,\ \ldots,\ X_n\}$ is crucial for accurate estimation using kernel density estimation. Through anonymisation, for example in the form of aggregation, these true coordinates are lost and replaced by the anonymised coordinates $W = \{W_1,\ W_2,\ \ldots,\ W_n\}$. Since only W is observable, while X remains unknown, systematic biases can arise in the estimation of spatial density. Spatial anonymisation can thus compromise the accuracy of the derived population densities and their covariates. This motivates the development of methods that explicitly take into account the measurement error introduced by anonymisation. The aim of these approaches is to reconstruct the original population density as accurately as possible and to reduce the biases caused by aggregation.

In addition to the bias that arises when using KDE on aggregated data, there are also limitations in the visualization of aggregated data. Aggregated data are often presented using choropleth maps, in which values are projected onto territorial units. This results in abrupt changes in values and colours at the boundaries. Such discontinuities can suggest false spatial patterns and cause kernel density estimates based on aggregated data to distort the actual structure. This results in the loss of fine spatial structures and makes it more difficult to interpret the distribution. The use of a naive kernel density estimator that ignores the aggregation process by replacing the true coordinates X with the anonymised W (in equation (1)) can result in a density with hotspots around the center point that is far from the density of the unadulterated (true) data. This effect becomes more pronounced as the sample size increases, since bandwidth estimators tend to become smaller as the sample size grows. This results in estimates with higher density at the center points and lower values between them.

Anyhow, choropleth maps manage to clearly illustrate the anonymisation process. Hence, assuming that the anonymisation process of X is known, is by no means unrealistic. With this knowledgement we are able to formulate a measurement error model $\pi(W|X)$. According to Bayes' theorem, $\pi(X \mid W) \propto \pi(W \mid X)\,\pi(X)$.

The first term describes the anonymisation process of X. For aggregation, the measurement error model $\pi(W|X)$ is defined as the product of Dirac distributions $\pi(W|X) = \prod_{i=1}^{n} \pi(W_i|X_i)$ with

$$\pi(W|X) = \begin{cases} 1 & \text{for } X_i \in P_a \\ 0 & \text{else,} \end{cases} \tag{5}$$

where $P_a,\ a = 1, \ldots, A$ denotes A aggregation areas over the area of interest, which are generally polygonal or

can take on a rectangular shape during the rounding process. Hence, the conditional density $\pi(W|X)$ describes the anonymisation mechanism, i.e., the probability distribution that governs how the exact coordinates X are converted into the observed anonymised coordinates W.

However, the second factor poses problems. Since the exact geocoordinates X are unknown, the density of X is also unknown. There is the possibility of an iterative procedure in which an estimate can be made by drawing a pseudo-sample of X_i from the distribution $\pi(X_i \mid W_i)$ for each i. Specifically, this means that a rough estimate of $\pi(X)$ is calculated based on the observed W. After that, the simulation of X from $\pi(X \mid W)$ and the updating of $\pi(X)$ alternate until a convergent state is reached. The kernel heaping algorithm introduced by Groß et al. (2016) uses this iterative principle to correct distortions in aggregated data. For aggregated data, density values using discretised pseudo-coordinates are determined by repeated kernel density estimation and sampling. This mitigates the artificial clusters at centers that arise with the naive method and enables supra-regional clustering, which is why the algorithm can be understood as an error correction method for reconstructing the underlying density. The pseudo-algorithm can be found in Groß et al. (2016).

3.3 Extension to regional proportions

In addition to aggregating geocoordinates, these are often set in relation to a subpopulation with specific characteristics by calculating the proportion of people with a certain characteristic within an area. The number of geolocations within an area is normalised by a second variable. The value for an area refers to the shares of the number of people with a specific characteristic to the total number in the area. The choropleth maps with percentage figures are intended to show in which regions people with a specific characteristic are more likely or less likely to be represented. The measurement error model and the resulting iterative procedure described can also be used to determine regionally smoothed percentages, if aggregated data are observed.

Let f_P be the density of the total population with n_P data points and f_C the density of the population with a specific characteristic with n_C data points. The expected number of people at evaluation grid point x_g with grid cell size $\Delta_1 \times \Delta_2$ is $n_P f_P(x_g)\Delta_1\Delta_2$. Similarly, the expected number of people with a specific characteristic is $n_C f_C(x_g)\Delta_1\Delta_2$. Therefore,

$$r(x_g) = \frac{n_C f_C(x_g)}{n_P f_P(x_g)} \tag{6}$$

represents the local proportion of individuals with a certain characteristic, which is corrected by the mean value across the population n_C/n_P. A non-parametric estimator of local proportions $r(\cdot)$ is the Nadaraya-Watson estimator $\hat{r}_{NW}(\cdot)$, see Härdle (1990) for details. The estimator can be represented as the ratio of two kernel density estimates, where a binary variable $Z_i \in \{0,\ 1\},\ i = 1,\ \ldots,\ n_P$ must be used. $Z_i = 1$ if data point X_i fulfills the characteristic and 0 otherwise. This results in the Nadaraya-Watson estimator $\hat{r}_{NW}(\cdot)$ for proportion estimates

$$\begin{aligned}\hat{r}_{\mathrm{NW}}(x_g) =& \frac{\frac{1}{n_P|H|}\sum_{i=1}^{n_P} Z_i K\left(H^{-1}(x_g - X_i)\right)}{\frac{1}{n_P|H|}\sum_{i=1}^{n_P} K\left(H^{-1}(x_g - X_i)\right)} \\ =& \frac{n_C \hat{f}_C(x_g)}{n_P \hat{f}_P(x_g)},\end{aligned} \tag{7}$$

where the estimates of the densities by kernel density estimation replace the true densities from equation (6). It is important that the same bandwidth H is used for the density of all geocoordinates and the density of those that exhibit the characteristic. Since the number of people with a characteristic is smaller than the total population, it makes sense to determine the smoothing factor based on the former data, as this is usually slightly larger.

The measurement error model accounting for aggregation from Section 3.2 must consequently be adapted. We

now introduce the resulting iterative algorithm for proportion data, for which the region of interest needs to be discretized. The evaluation grid resulting from the discretization can be described by x_g, $g = 1, \ldots, G$, representing the geocoordinates of the G evaluation points, where Δ_1 and Δ_2 denote the distance between two grid points in the longitude and latitude directions, respectively. For each grid point, it must be determined in which of the A areas it lies. This divides the grid points into A subsets $\mathcal{G}_a = \{x_g \mid g = 1, \ldots, G;\ x_g \in P_a\}$, where $a = 1, \ldots, A$ corresponds to the areas. The division is disjoint, i.e., the set of grid points $\mathcal{G} = \cup_{a=1}^{A} \mathcal{G}_a$. The centers of the area a are the obtained anonymised points which are denoted by W_a. For area a, there are $N_{a,P}$ observed values from the total population and $N_{a,C}$ from those with a specific characteristic, where $\sum_{a=1}^{A} N_{a,P} = n_P$ and $\sum_{a=1}^{A} N_{a,C} = n_C$.

For a detailed overview, see the pseudo-algorithm 1.

Algorithm 1 Pseudocode: Proportion estimation treating aggregation as a measurement error.

1: Given: aggregated data W_a, their associated polygons P_a and the observed integer values $N_{a,P}$ and $N_{a,C}$, $a = 1, \ldots, A$, number B of burn-in and number S of sampling iterations
2: Specification: evaluation grid $\mathcal{G}$ and resulting set of potential pseudo-samples $\mathcal{G}^a$, $a = 1, \ldots, A$.
3: Calculation: naive kernel density estimation $f_C^{(0)}$ and $f_P^{(0)}$ based on aggregated data W_a and $N_{a,C}$ or $N_{a,P}$ with large bandwidth
4: **for** t in $1 : (B + S)$ **do**
5: Drawing of $N_{a,P}$ pseudo-sample points from the evaluation grid $\mathcal{G}_a$, proportional to $f_P^{(t-1)}$, for $a = 1, \ldots, A$. The total sample $s_P^{(t)}$ comprises n_P pseudo-samples.
6: Draw $N_{a,C}$ pseudo-sample points from the total sample $s_P^{(t)}$, taking into account the aggregation limits, proportional to $r^{(t-1)}$, for $a = 1, \ldots, A$. The total sample $s_C^{(t)}$ comprises n_C pseudo-samples.
7: Calculation of kernel density estimates $f_P^{(t)}$ and $f_C^{(t)}$ based on $s_P^{(t)}$ and $s_C^{(t)}$, respectively. The bandwidth was calculated based on $s_C^{(t)}$ according to Wand and C. Jones (1994).
8: Calculate the local proportions $\hat{r}^{(t)} = \frac{N_C f_C^{(t)}}{N_P f_P^{(t)}}$
9: **end for**
10: Estimated proportions $\hat{r} = \frac{1}{S} \sum_{t=1} \hat{r}^{(t+B)}$.

In each iteration, a set of pseudo-samples is drawn that is intended to imitate the population density to be estimated as closely as possible. According to Bayes' theorem $\pi(X \mid W) \propto \pi(W \mid X)\,\pi(X)$ the new sample is drawn proportionally to both the aggregation process (see equation (5)) and the previous estimate of the density. The number of pseudo-samples drawn in each area corresponds exactly to the number of observations previously aggregated for that area. Furthermore, the pseudo-samples must be assign having the characteristic or not. For this, the same number of pseudo-samples is selected as people having this characteristic have been aggregated in this area. Furthermore, this assignment is performed according to its proportion values from the pervious iteration. The subpopulation (subset of the pseudo-samples) is drawn from the population (pseudo-samples) randomly according to the proportions obtained from the previous iteration. This proportional selection procedure leads to reinforcement of existing effects. Within the framework of the measurement error model, the population and the subpopulation drawn from the set of evaluation points within an aggregate, but are limited in terms of quantity and space. These limitations mean that existing clusters are systematically reinforced by the proportional sampling procedure, while at the same time the structural limitations have a stabilising effect, so that the resulting accumulations remain controlled.

The algorithm goes back to the work of Groß et al. (2016). The implementation of the described algorithms was published in the R package *Kernelheaping*, and Gril et al. (2025) describe the detailed use of the package in the vignette. In the sections that follow, the implementation uses the default settings.

4 Application to (high) income tax payers in Berlin

For scientist having access to geolocation data in official statistics offers a rare opportunity for profound analysis, a resource not typically available in this research field. For the first time, the 2019 cohort of the income tax data in Berlin equipped with geocoordinates is available, for a detailed description of the data, see section 4.1. The focus of the following pages lies on an efficient regional analysis of high earners. By showing densities of income tax payers (ITP) and high income tax payers (HITP - defined in Section 4.1) as frequency data and as smoothed densities (Section 4.2), we will both, motivate the use of smoothed maps and encourage the use of proportions for a meaningful analysis, see Section 4.3. Furthermore, since wage and income tax statistics are highly sensitive with regard to data protection, anonymisation methods such as aggregation are applied to the geocoordinates. Changing the geocoordinates, however, results in information loss, and the true proportions of HITP across Berlin cannot be preserved. In Section 4.3.1 and 4.3.2, the measurement error model introduced in Section 3.3 is applied in order to obtain smoothed proportions on anonymized information as accurately as possible and quantifying the error caused by anonymization.

4.1 Description of the data

For the first time, the statistical offices in Germany have access to addresses in combination with wage and income tax statistics of the cohort of 2019. Wage and income tax statistics are compiled annually in Germany and basically include all information contained in wage tax returns, such as details of income and types of income. In addition, demographic, social and economic structural characteristics such as place of residence, year of birth, gender, religious affiliation, child allowances and child benefit, economic sector or type of liberal profession, type of tax liability, tax class and type of assessment are recorded. In the underlying analysis, we had access to the address data, taxable income and type of assessment of the taxpayers, see the technical report of the wage and income tax statistics Statistisches Bundesamt (Destatis) (2021) for details.

The address data must be converted into coordinates for meaningful regional analysis. The internal geocoder of the Statistics Office in Berlin-Brandenburg was used for geocoding. However, for 12% of the addresses, it was not possible to make a clear assignment within Berlin, either because the addresses could not be found or because of relocation, whereby the new address was recorded outside Berlin.

The analysis is based on taxable income, which is the basis for assessing income tax. The total income – a distinction is made between seven types of income – is reduced by certain allowances and reliefs, resulting in the taxable income.

Furthermore, a distinction is made between four types of assessment, namely individual assessment of spouses/partners, joint assessment, other assessment and widow splitting. Jointly assessed persons are counted as one taxpayer but are to be treated as two tax cases, since two individuals are involved. All other taxpayer represents one tax case. Since we have no further individual information and the aim is to analyse the spatial distribution of tax cases, jointly assessed persons are treated as two separate tax cases and their joint income is divided equally between them. Of the approximately 1,925,200 taxpayers in the data, around 394,000 are jointly assessed (20%), which, after deducting the 12% of data to which no address could be assigned, results in approximately 2,042,400 tax cases for the 2019 tax year in Berlin. Those tax cases that are recorded in the wage and income tax statistics and used for analysis are referred to as income tax payers (ITP).

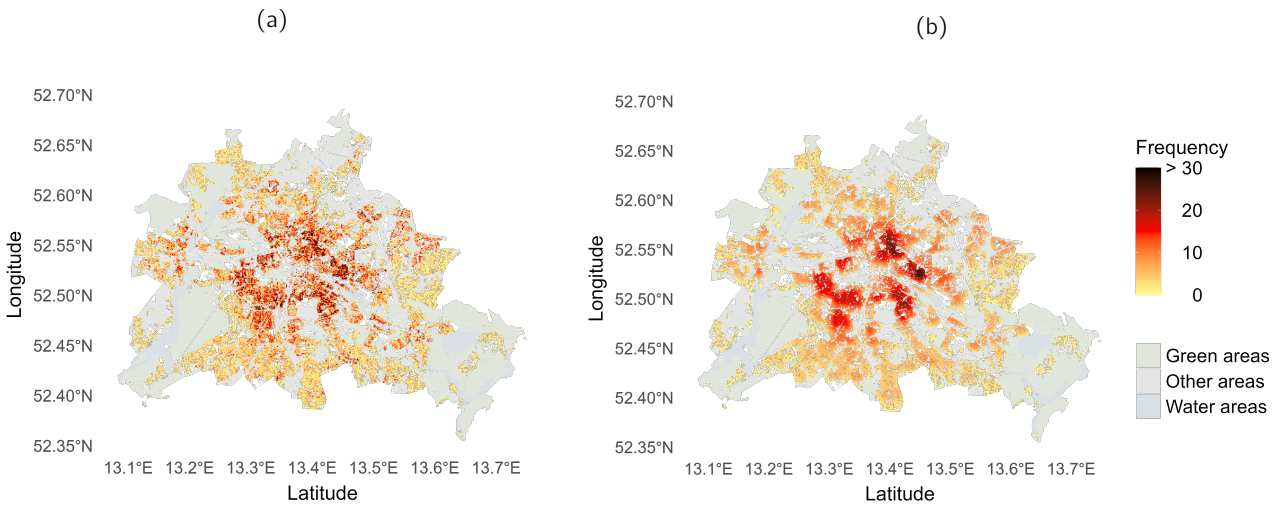
For a more detailed explanation of how taxable income is calculated and a detailed analysis of how taxable income can be split for jointly assessed persons, we refer to Appendix A.

A more detailed spatial analysis is carried out for ITPs with high incomes. For this purpose, the 90th percentile of the annual income distribution was calculated. The threshold value in our analysis is approximately € 54,100. Individuals with a taxable income above this threshold are referred to as high-income taxpayers (HITP). First, the spatial distribution of ITP and HITP and the visualisation potential through smoothed estimates using Kernel density estimation (KDE) are discussed.

4.2 Regional density of income taxpayers and high income taxpayers

First, we examine the regional distribution of ITP and HITP and temporarily set aside the issue of anonymization. Although the wage and income tax statistics data is available with exact addresses, i.e., with geocoordinates in the form of longitude and latitude in the metric LAEA coordinate system, cf. Federal Agency for Cartography and Geodesy in Germany (2019), the question of suitable visualisation arises. As already argued by Baddeley et al. (2015), the visualization of individual data points is meaningless in this context. In order to enable the finest possible spatial evaluation and at the same time to highlight the problem of unsmoothed representations, the Berlin city area is discretised on the basis of a grid with a resolution of 100×100 meters. Based on exact geocoordinates, a 10% sample of the ITP is assigned to the corresponding grid cells (see Figure 4.1a). However, representing frequeny data in grid cells presents several challenges. On the one hand, grid cells with few or no observations are created, which are particularly noticeable in areas with uninhabitable infrastructure. On the other hand, there are numerous cells with a large number of observations, especially in densely populated residential areas with apartment complexes. These cannot be represented in one legend, which is why grid cells representing more than 30 people have been assigned a uniform color (black). Water area, green spaces, and other areas (such as hospitals, schools, etc.) are shown in very light shades of blue, green, and gray, respectively. Other uninhabited or uninhabitable areas – streets, for example – are shown in white. In all subsequent graphics, grid cells that cannot be displayed for data-protection reasons are also shown in white. As a result, they cannot be distinguished from uninhabited areas, but no additional attention is drawn to them. Our focus is the visualization of spatial patterns. Of the total of 171,748 grid cells covering the Berlin city area, with a north-south extension of approximately 38 kilometers and an east-west extension of approximately 45 kilometers, 41,156 are inhabited. The grid cells were created based on the northernmost, southernmost, westernmost, and easternmost points of Berlin.

Figure 4.1: Illustration of the (a) frequencies and (b) KDE of a 10 % subsample of the ITP on a 100×100 m grid



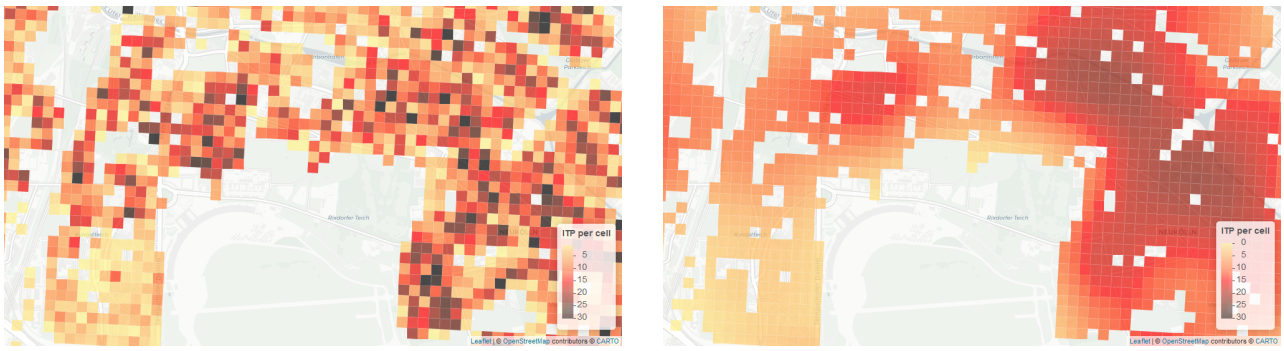
In addition, the exact geographical coordinates of the 10% sample of the ITP are used to calculate a smoothed

representation using KDE, shown in Figure 4.1b. The evaluation is again carried out on those grid points that were inhabited when the Berlin city area was discretised with a resolution of 100×100 meter. A boundary correction must be used for the calculation, where probability mass lying outside the inhabited areas is reflected back, cf. Section 3.1. It is clear that the smoothed map does not show the fine details of the infrastructure, but rather captures large-scale structures. It reveals the regions in which comparatively many or few ITPs live, thereby highlighting overarching patterns. These large-scale areas reflect clusters and show the spatial concentration of the ITPs.

When zoomed in, the map reveals the strong smoothing effect of the kernel density estimation compared to the evaluation of the original data, see Figure 4.2. The unsmoothed map mostly shows abrupt differences between neighboring cells. This is mainly due to the fact that the rigid, straight-line grid is applied to a dynamic and heterogeneous cityscape. Some cells cover several apartment buildings, while others overlap with roads or parking areas, leading to abrupt changes in the frequency values. These differences are sensitive to the selected grid size and the location of the cell centers. Shifting the grid by a few dozen meters can produce significantly different results. In contrast, the map smoothed with KDE provides a higher-level, continuous view that clearly shows whether there are many or few ITPs concentrated in an area.
In summary, it can be stated that smoothed representations should be preferred over frequency data on grid cells. Although only inhabited areas are shown, the smoothed visualization clearly conveys the key structural patterns.

Figure 4.2: Zooming in on the (a) frequencies and (b) KDE of a 10 % subsample of the ITP on a 100×100 m

(a) (b)

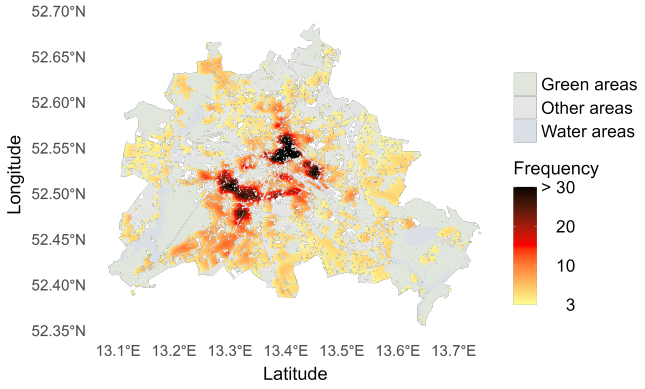


In addition to analyzing the data's visualization potential, the focus of this work is on the regional analysis of HITP. Hereby, Figure 4.3 shows the regional distribution of HITP using KDE. All HITPs are included in the estimation to ensure that the legends are comparable with those in Figure 4.1. Grid cells with less than 4 inhabitants are not displayed. The evaluation is performed on the same 100×100 meter grid. As expected, it can be seen that HITPs occur more frequently in areas with many ITPs. However, a closer look reveals a heterogeneous picture. There are clear peaks in new development areas around the main railway station in the center of Berlin and along the Grunewald forest in the south-west, while lower values can be observed in districts characterized by prefabricated buildings, such as Marzahn-Hellersdorf in the east. Since absolute figures hardly reveal in which areas HITP are strongly represented relative to ITP, the following section focuses on proportion representations.

4.3 Regional shares of high income taxpayers

The main focus of our analysis is to create an efficient and reliable representation of high-income taxpayers. Representation of proportions make it possible to normalise absolute frequencies to the underlying population or total population. This allows regional differences to be compared, regardless of the size or population density of the areas. Such maps facilitate the identification of relative centres of gravity and clusters and prevent densely populated areas from dominating the interpretation. First, we focus again on the visualization, but due to the high sensitivity

Figure 4.3: The KDE of exact coordinates of all HITP on a 100×100 m grid



of the data, we will address its anonymization later.

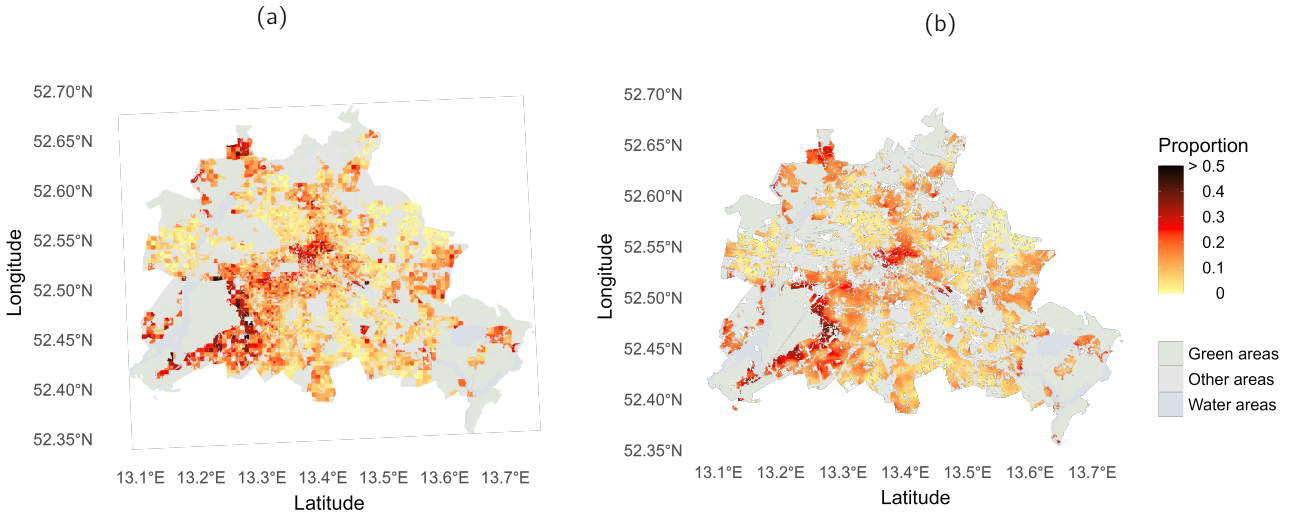
Figure 4.4a shows the true regional proportions based on the frequency data of HITPs relative to ITPs using different sized grid cells and a 10% sample of the ITPs. In order to be able to display the graphic reliably, areas with fewer than 30 ITPs were iteratively combined with their respective three neighbouring cells using the quadtree method until this threshold was reached. The quadtree method enables adaptive aggregation of spatial data by recursively summarizing areas into quadrants until a desired anonymity threshold is reached, thus preventing conclusions about individuals, cf. Strobl (2004) and Behnisch et al. (2013). In addition, cells in which more than 50% belong to the group of HITPs are displayed uniformly (black). Again, the erratic representation makes it difficult to identify local clusters, which is partly due to infrastructure or the heterogeneous population structure. Nevertheless, this map shows that the proportion of HITPs is particularly high in attractive residential areas, for example along the Grunwald and at the lakes in the south-west of Berlin.

The smoothed representation using the Nadaraya-Watson estimator, which sets the KDE of the HITPs in relation to the KDE of the ITPs, makes it easier to identify large-scale patterns, see Figure 4.4b. To ensure a secure visualization, a 75% sample of the ITPs was used, and additionally, 100×100 m grid cells with fewer than five ITPs were not displayed. For this purpose, the kernel density estimates of the ITPs and HITPs were calculated with a uniform bandwidth and normalised by a factor of 0.1, compare equation (6). The normalisation factor is based on the definition of a HITP. Smoothing attenuates extreme local fluctuations, making disproportionate concentrations of HITPs in larger contiguous areas more prominent.

From these maps, it is again evident that smoothed representation enables to display the essential spatial structure and are far less security-sensitive, since the smoothing procedure alone makes re-identification more difficult. To ensure that the re-identification of individual persons is only possible with a disproportionate level of effort, both sampling and the omission of information for cells with very small populations were employed in the map representations. Furthermore, it is noteworthy that not only the case where a grid cell has a proportion of 1, i.e., all ITPs are HITPs, is sensitive, but also a proportion of 0. Even knowing that no HITP lives in a grid cell reveals information. We were able to circumvent this issue by sampling, as even cells with a proportion of 0 retain a residual probability that an HITP resides there.

Due to their high informational content, income tax data are among the most sensitive official data sets. Accordingly, anonymisation strategies are becoming the focus of attention. The central focus of the present analysis is the question of to what extent an efficient graphical analysis is possible despite the aggregation of exact

Figure 4.4: Illustration of the (a) proportions of HITPs relative to ITPs on grid cells of different sizes obtained by the Quadtree method as well as (b) smoothed proportions of HITPs relative to ITPs using Nadaraya Watson estimator on a 100 × 100 m grid



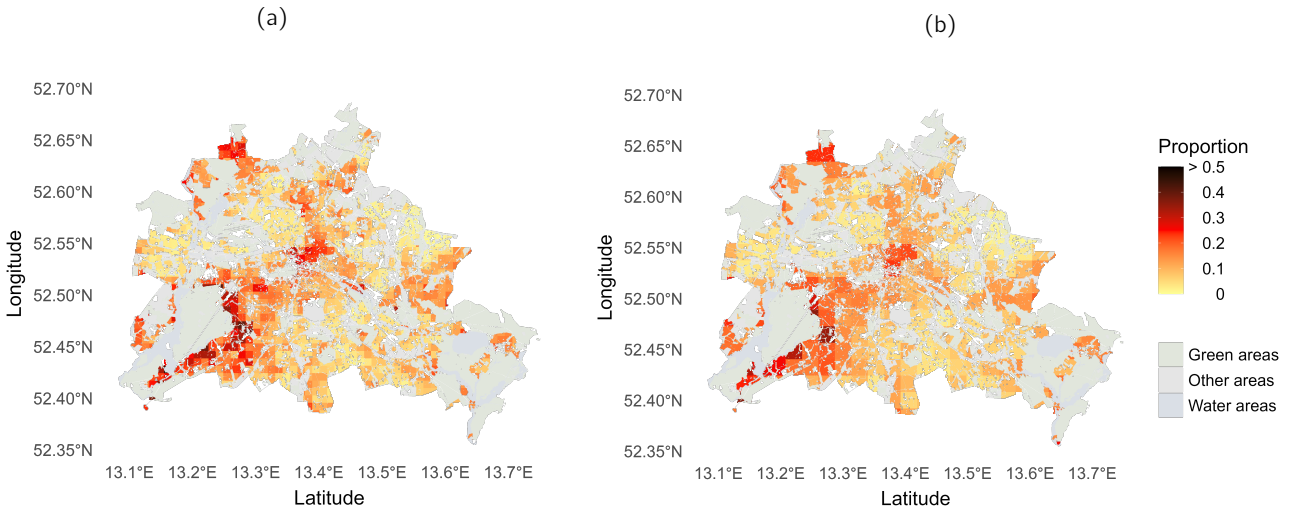
geocoordinates into larger spatial units through the application of the measurement error model. First, a systematic enlargement of the spatial aggregation units is examined. In addition, an application scenario is constructed linking census data to illustrate the practical implications of aggregation under realistic conditions.

4.3.1 Influence of spatial aggregation levels on anonymisation effects in regional shares of high income taxpayers

By analyzing the visualization of the regional shares based on exact geocoordinates, it is now clear that a smoothed representation of the data is preferable. However, statistical offices generally do not publish these data, and certainly not as machine-readable maps, meaning that users do not have access to smoothed proportion values at grid level. In contrast to machine-readable data, reconstructing the values from the map would require considerable effort. Therefore, only aggregated data are usually published in official statistics. With the publication of aggregated data, the question arises whether alternative representations, in addition to choropleth maps, can provide a more accurate depiction, since the distortions associated with choropleth maps are amplified, particularly at high spatial granularity. By treating aggregation as a known measurement error, pseudo-samples are drawn, which enables a smoothed representation of the proportion values.

In order to be able to perform controlled calculations based on different aggregation levels of address-specific data, the exact geocoordinates of the ITPs and HITPs are first rounded to rectangles of different sizes. Aggregation by rounding creates uniform areas of equal size. However, these areas contain widely varying population figures. Edge lengths between 200 and 2000 meters are used to illustrate the anonymization effects of aggregation. The number of ITPs and HITPs in each aggregate is now known, and thus the proportion values for the entire aggregate. As an example, Figure 4.5 shows the proportion values of the aggregation rectangles of size 800 × 800 and 1800 × 1800 meter. The anonymisation of data through rounding is accompanied by significant limitations in the interpretability of small-scale structures. With increasing aggregation, the accuracy of the data decreases, resulting in a natural tendency towards a globally representative share of approximately 10% in the aggregated areas. Furthermore, representations of aggregated data are often characterised by pronounced jumps along the rounding boundaries. Since the visualizations are based on exact geocoordinates, a 90% sample was drawn for secure graphical representation, and all aggregates with fewer than 100 ITPs were not displayed.

Figure 4.5: Illustration of the aggregated proportions for (a) 800 × 800 and (b) 1800 × 1800 meter aggregation rectangles



Using information on the position and shape of the aggregates, as well as the number of ITPs and HITPs assigned to them, the measurement error model accounting for aggregation (Algorithm 1) is applied to produce a smooth representation of the proportions. Note that in contrast to Figure 4.5, the evaluation of the method is based on the full dataset, not a sample. The procedure is intended to enable a smooth representation of the data by iteratively drawing pseudo-samples that approximately represent the places of residence of ITPs, taking into account the aggregation as a measurement error and using KDE. In addition, depending on the frequency of HITPs in the respective aggregate, pseudo-sample points representing ITPs are reported as HITPs in order to enable the calculation of regional proportions. The smooth representations, see Figure 4.6, are based on the algorithm's estimation of regional proportions, with the initial data for the method being the proportions rounded to an 800 × 800 or 1800 × 1800 meter rectangles.

The graphical representation of the proportions contributes significantly to improving interpretability and enables the identification of local clusters. Since the algorithm aims to produce smoothed proportions, the estimated proportions based on aggregated data (see Figure 4.5) must be compared with each smoothed proportion using the original coordinates (see Figure 4.4). It is clear that the loss of information increases as the size of the aggregates increases.

These observations can also be quantified empirically using the mean absolute deviation (MAD), a global error measure. Let $r_1(\cdot)$ and $r_2(\cdot)$ be two proportion functions. Then the mean absolute deviation between the two proportions on the evaluation grid $\mathcal{G}$ over the area of interest under consideration is given by

$$MAD(r_1, r_2) = \frac{1}{G} \sum_{g=1}^{G} |r_1(x_g) - r_2(x_g)|. \tag{8}$$

The MAD between the smoothed proportions based on the exact geocoordinates and shares obtained by using the frequencies of the HITPs and the ITPs per cell is 2%. The evaluation is carried out on a 100 × 100 meter grid over the urban area, whereby only inhabited areas are used for the evaluation. The sudden changes in the proportions, which are due to infrastructural and other spatial circumstances, make it necessary to carry out all comparisons on the basis of the smoothed proportions derived from the exact coordinates.

Figure 4.7 shows the MAD values (vertical axis) for different rounding values (horizontal axis). The regional proportions determined using the Nadaraya-Watson estimator based on the exact geocoordinates serve as a comparison value.

Figure 4.6: Illustration of the smoothed proportions using the measurement error model, which used the frequencies of ITP and HITP aggregated on (a) 800 × 800 and (b) 1800 × 1800 meter aggregation rectangles

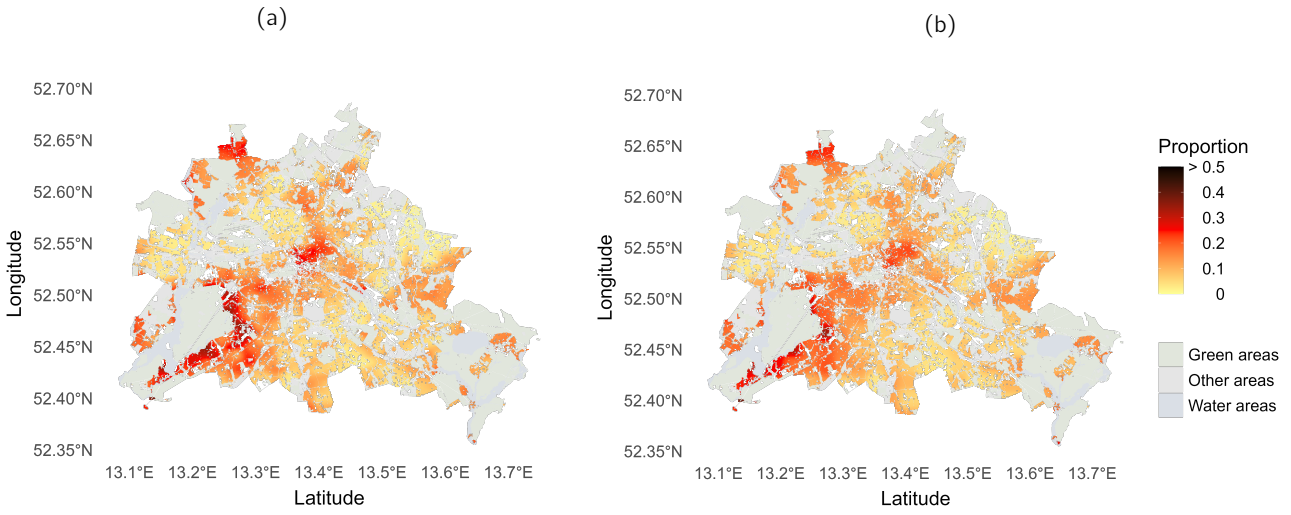
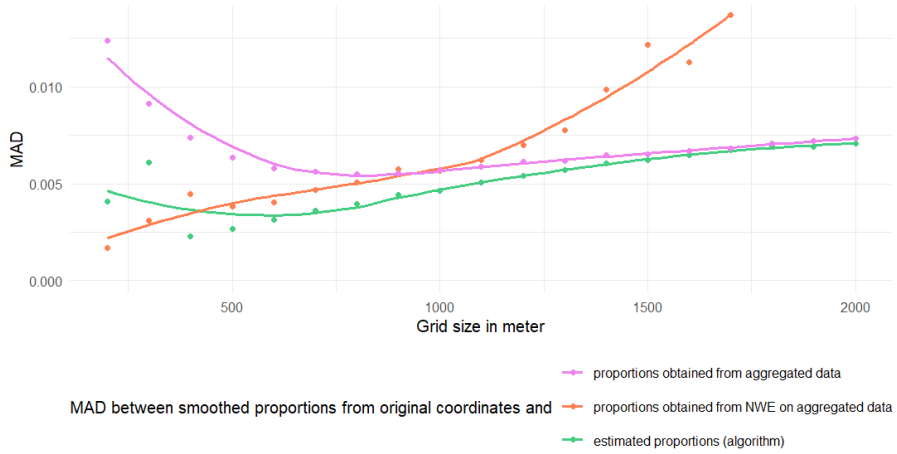


Figure 4.7: MAD comparing smoothed proportions form original coordinates with aggregated data (violet), naive smoothed proportions using anonymised coordinates (red) and smoothed proportions obtain from the measurement error model (green) cross varying aggregation rectangle sizes



For the aggregated proportions, it is on the one hand possible to assign the proportion value to the entire aggregate. On the other hand, it is also possible to use the anonymized coordinates W and calculate a smoothed proportion estimate based on them. The former is shown in the violet-colored curves, the latter in the orange-colored curves.

Regarding the first approach, smoothed proportion values based on the exact coordinates are compared with the respective aggregated proportions for different rounding values (violet-colored curves). The higher the rounding value of the aggregation, the closer are the proportions in the aggregation areas to the global proportion of 10%. As a result, the MAD stabilizes as the rounding values of the aggregation increase.

Furthermore, a naive smoothed estimate of the proportions can be calculated based on the aggregated geocoordinates W (orange-colored curve). Aggregation leads to biased estimates, especially for high rounding values. Aggregating the exact geocoordinates effectively reduces the bandwidth for higher rounding values, causing a concentration on the aggregation points and resulting in increasingly distorted estimated densities and proportions. Because the results could no longer be meaningfully represented due to high deviations, the orange curve had to be truncated at a rounding value of 1700. Therefore, a naive estimate based on W is not recommended.

The green curve shows the MAD value of the comparison between the smoothed share values based on the original coordinates and the share values estimated using the measurement error model, taking the number values of the ITPs and HITPs per aggregation area into account. It can be seen here that the estimates are not useful for small aggregation values, but for larger aggregation values they lead to a stable reduction in the MAD. Since the evaluation is performed on a 100×100 meter grid, there are simply too few alternative points available in each aggregate for small rounding values up to 500 meters to improve the estimate. Hence, the potential of the measurement error model depends on the number of feasible pseudo-coordinates. If too few such candidates fall within the aggregation area, the method cannot fully realize its effectiveness. The higher the rounding values is, the greater is the loss of information, and thus only a slight improvement in the MAD is possible compared to considering the aggregates.

4.3.2 Empirical scenario analysis linking census information

In addition to the systematic analysis of aggregation areas of varying sizes, it is common practice in official statistics to provide data at clearly defined administrative levels. Therefore, the aggregation and the measurement error model should be tested on administrative boundaries. Spatial administrative boundaries for Berlin can be found as polygons in the Open Data Portal, see Senatsverwaltung Berlin (2026), and are distinguished by their level of detail or loss of information. The more polygons there are, the more detailed is the aggregation and the lower is the loss of information compared to the original coordinates.
There are twelve Berlin districts (BEZ), 58 forecast areas (PRG), 96 subdistricts (ORT), 138 district regions (BZR), 193 postcodes (PLZ) and 542 planning areas (PLR). The structure of the polygons has grown naturally in some cases, for example along river boundaries, but includes areas of varying sizes and widely varying population figures.

It is further assumed that frequency data for ITPs and HITPs are available at the lowest level, i.e., the planning areas. Using the exact polygon geometries, the measurement error model can then be applied. Iteratively, pseudo-coordinates are drawn that mimic the original coordinates of the ITPs respecting the number of ITPs per aggregation area and the previous density of the ITPs. From the set of pseudo-samples representing ITPs, a subset is classified as HITPs according to the number of HITPs in the corresponding aggregation unit and the proportions of HITPs relative to ITPs of the previous iteration. By proportional sampling, areas with a high HITPs share are reinforced, but respecting the frequency values of HITPs and ITPs per aggregation area has stabilizing effects. Through kernel smoothing, the result is a smoothed representation of proportion values across the city.

The prerequisite for drawing pseudo-coordinates for the HITPs is the availability of pseudo-coordinates for the ITP. Due to the aggregation of the ITPs, their regional distribution must be estimated, which is therefore subject to substantial bias. Consequently, it is a question whether there exists a suitable proxy for the regional distribution of the ITPs. In this case, the distribution of the ITPs would not need to be estimated; instead, pseudo-coordinates could be drawn proportionally to the proxy while still accounting for the aggregation. This means that the information about the position of ITPs is borrowed by the proxy, while accounting for the frequency values of ITPs per aggregation area. In step 5 of Algorithm 1, the proxy is used instead of the population density from the previous iteration, $f_P^{(t-1)}$. The determination of the subpopulation is carried out analogously to the previous procedure. The subpopulation respresenting HITPs is drawn from the set of pseudosamples mimicing ITPs proportionally to the proportion values from the preceding iteration. Only the iterative determination of the ITPs is linked to the proxy.

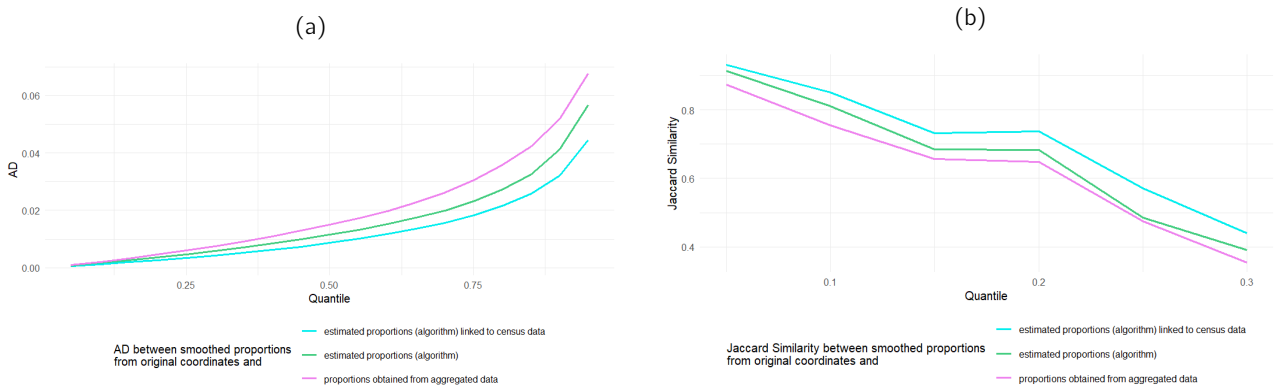
Among other interesting characteristics, the 2022 German Census includes the determination of the population. Using the Cell-Key method, the data were anonymized and can be represented on a 100×100 meter grid. The Cell-Key method is a data-perturbation disclosure control technique that introduces small, random but fixed mod-

ifications to aggregated frequencies, preventing direct inference about original values. The perturbation is applied consistently across tables and queries, ensuring logical coherence while minimizing information loss, cf. Fraser and Wooten (2005) and Hundepool et al. (2018). This data is publicly available at Statistisches Bundesamt (Destatis) (2024) website. Restricted to the Berlin city area, the population data serve as our proxy. It should be noted that the data do not reflect the regional distribution of the ITPs, which is a subpopulation of the total population. The residents of Berlin form a group of approximately 3.59 million people, of which 2.04 million are ITPs. Moreover, the temporal discrepancy should be noted.
The global measure RMISE (see Appendix B.1 equation (10) for details), which describes the root of the squared error between two densities at each evaluation point, is very low (in the order of 10^{-6}). Furthermore, no systematic regional deviation patterns are apparent. Accordingly, Census data can serve as a plausible proxy for the density of the ITP.

The measurement error model as an iterative procedure (Algorithm 1), as well as the integration of this procedure with the census data, makes it possible to obtain estimates of smoothed proportion values from the aggregated data at the level of planning regions (PLR). Since the graphical advantages of this smoothing have already been discussed in detail in the previous sections, we focus here on error measures.

For this purpose, Figure 4.8a shows the quantiles of the absolute deviation (QAD), with quantile ranges from 0.05 to 0.95 in steps of 0.05. Note that the evaluation took place on a 100×100 meter grid. In contrast to Section 4.3.1, which focused on the mean absolute deviation, our objective here was to conduct a more in-depth analysis. As a reference, the regionally smoothed proportions of HITPs relative to ITPs based on the exact geocoordinates are used. The purple curve represents the quantile-based absolute deviations in this range compared to the aggregated data interpreted as chorpleth map. The smoothed proportion estimated by applying the measurement error model accounting for aggregation are shown in the green curve. The bias introduced by aggregation can be reduced. In addition, the proxy for the regional distribution of the ITPs is employed. Linking the algorithm with census data allows for a further reduction of the absolute deviations with respect to the quantiles considered. The blue curve shows the corresponding QAD values. The results indicate a pronounced increase in quantile-based absolute deviations (QAD) in the upper quantile ranges.

Figure 4.8: (a) Absolute deviance (AD) for given quantiles and (b) Jaccard Similarity for different hotspot areas comparing smoothed proportions from original coordinates with aggregated data (violet) and smoothed proportions obtain from the measurement error model with (cyan) and without (green) using census information



Of particular interest is whether areas whose HITPs proportions exceed a given threshold of $p\%$ can be reliably identified in both the estimated and the true smoothed proportions. For this purpose, the Jaccard similarity is used as a local measure. Let $r_1(\cdot)$ and $r_2(\cdot)$ again denote two proportion functions. The Jaccard similarity then measures the agreement of areas whose proportions exceed the threshold $p\%$. Let $U = \{x_g \in \mathcal{G} \mid r_1(x_g) > p\%\}$

and $V = \{x_g \in \mathcal{G} \mid r_2(x_g) > p\%\}$, then

$$J(U,\ V) = \frac{|U \cap V|}{|U \cup V|} \quad . \tag{9}$$

A value of $J(U,\ V) = 1$ indicates complete agreement, while $J(U,\ V) = 0$ indicates complete discrepancy.

In Figure 4.8b, the Jaccard similarity, evaluated on a 100×100 meter grid, is shown for thresholds from 5 to 30% in 5% increments. As in Figure 4.8a for the QAD values, the regionally smoothed proportions of HITPs relative to ITPs based on the original geocoordinates are used as a reference. The purple curve represents the comparison to the aggregated data, the green curve to the smoothed proportion estimates based on Algorithm 1, and the blue curve to the smoothed estimates from the algorithm linked with the census data. It is also evident here that local hotspots are better detected when the proposed method is applied to aggregated data. A clear improvement is again achieved through linkage with the census. The proxy allows the underlying population structure to be reflected and improves the results, even though it does not directly affect the determination of who is assigned as HITP.

5 Conclusion

Exact geocoordinates form the basis for nuanced spatial analyses. However, they usually cannot be represented as point data on maps due to confidentiality constraints and visualization problems. Geocoordinates can be aggregated into grid cells, but this makes the analysis of local clusters more difficult and could not be secure for publication. Here, kernel density estimation (KDE) offers smoothed representations that highlight spatial patterns and clusters while simultaneously providing privacy-friendly smoothing of sensitive information.

A distinction must be made between the cartographic representation and the final data transfer to the receiver, as location data in particular is highly sensitive information and subject to strict data protection regulations. The publication of machine-readable data requires different data protection measures than the provision of maps, where the effort required to precisely assign values to cells would be disproportionately high.

Nevertheless, also graphical representations require the implementation of data-protection measures. The use of exact official geocoordinates entails several challenges. Sensitive grid cells had to be excluded, and sampling procedures were employed. In this context, it is crucial not only to consider the information conveyed by the map but also to ensure that individuals cannot infer information about others based on their own knowledge. Sensitive information is not limited to a person being a HITP, the absence of this characteristic is equally informative. If it can be determined that all people in a cell are not HITPs, individual information is revealed, which must be protected. By relying on sampling and omitting cells with very few ITPs, we were able to prevent such cases.

In official statistics, the aggregation of data linked to geocoordinates is an established method for making data accessible while simultaneously making it impossible to launch potential attacks without disproportionate effort. However, once aggregated data has been published, the question arises whether considering known measurement errors enables alternative forms of representation besides choropleth maps. The problems associated with choropleth maps increase with increasing data granularity.

Building on the idea of treating anonymization as a measurement error, a Markov chain Monte Carlo approach was developed. Here, aggregation was treated as a measurement error under the assumption that the aggregation process is known. By representing the anonymized data as a choropleth map, the assumption of a known anonymization process is realistic. The iterative algorithm generates pseudo-geolocations intended to simulate the underlying distribution of the latent original locations. Furthermore, the algorithm was extended to allow for estimation of regional proportions. For this purpose, the characteristics were randomly assigned to the drawn

pseudo-sample points proportional to the results of the previous iteration.

By having access to exact geocoordinates, it was possible to systematically analyze the effects of aggregation. If aggregation areas are small, i.e., rounding into rectangles with a edge length below 500 m, the choropleth maps are noisy and large error occur. Additionally, as the evaluation was conducted on a 100×100 meter grid, the measurement error model could realize its intended effect, as the feasible set of pseudo-coordinates to sample from is too small. In this case, a naive estimate based on the aggregated geocoordinates remains relatively efficient. However, as the aggregation sizes increase, the naive estimate became severely biased. The information loss, measured by the mean absolute distance, also increased for the representation as choropleth map. Applying the measurement error model reduced the error. However, for large aggregation units, up to 2000 meters, the information loss was too large and the measurement error model reached its limit. In such cases, the smoothed representation and the aggregates tend to converge toward the global proportion and, in particular, that local clusters within an aggregation unit cannot be reconstructed.

In addition to rounding as an aggregation process, which results in quadratic aggregation areas, the data were aggregated to planning areas. Aggregation to administrative boundaries is characterized by heterogeneous forms and sizes. In addition to using administrative boundaries, we considered how the estimates could be improved. Since census data depict population distribution and ITP can be seen as a sample of them, this data served as a proxy for the regional distribution of the ITPs. By linking the census information with the measurement error model considering aggregation, a further improvement was achieved with regard to the global error measure QAD (quantiles of the absolute deviation) and the local error measure Jaccard similarity. The improved estimation of the density of the ITPs allowed the regional proportions to better align with the smooth proportions of the true coordinates. It is important to note that the characterization of the pseudo-sample is not directly affected by its property (in this case, being a HITP).

In summary, access to official geocoordinates enables both a detailed analysis of map visualization and a systematic investigation of the effects of spatial aggregation. Smoothing-based visualizations capture essential structural features and allow for the correction of uninhabited areas, whereas frequency maps, while able to reproduce the data more precisely, are consequently noisier. By using aggregated and consequently highly distorted data as input, the measurement error model produces smooth maps, while clearly revealing the extent of information loss. Global and local error metrics confirm the empirical effectiveness of the model.

Acknowledgment

The work of Lorena Gril was supported by the joint project AnigeD under grant number 16KISA097. We thank the Statistical Office of Berlin-Brandenburg for supporting this project, providing the database, and granting access to the data. We also thank Martin Möhler (Federal Statistical Office, Destatis) for his critical review and constructive comments.

A Calculation scheme of taxable income in Germany

The wage and income tax statistics are an annual official register in Germany, conducted for each calendar year (January 1st to December 31st) since 2012. A detailed breakdown of how taxable income is calculated is provided in Table A.1.

Married couples filing jointly are treated as a single taxpayer in the German wage and income tax statistics. For our analysis, however, they need to be treated as two tax cases, since two individuals are associated with

Calculation of Taxable Income
Income: Agriculture and forestry, business enterprise, self-employment and employment, capital assets, rental/leasing, other income
= Total income
Deductions: Old-age relief amount, single-parent relief amount, relief amount for farmers/foresters
= Total adjusted income
Further deductions: Special expenses, extraordinary burdens, pension contributions, tax incentives (housing/monuments), loss deduction
= Income
Deductions: Child allowance, hardship allowances
= Taxable income

Table A.1: Calculation scheme of *taxable income*

one taxpayer. We do not have separate data on the taxable income of the two spouses. To make the analysis interpretable, the taxable income of these jointly assessed taxpayers must be split between the two cases. In addition to the equal split used for the analysis in Section 4.3, alternative splitting ratios of the joint income were tested. A one-third split represents a more extreme case, whereas the 41/59 and 47/53 splits correspond to the unadjusted and adjusted gender pay gap of 18% in 2020 and 6% in 2025, respectively. For a detailed overview see Statistisches Bundesamt (Destatis) (2025) website.

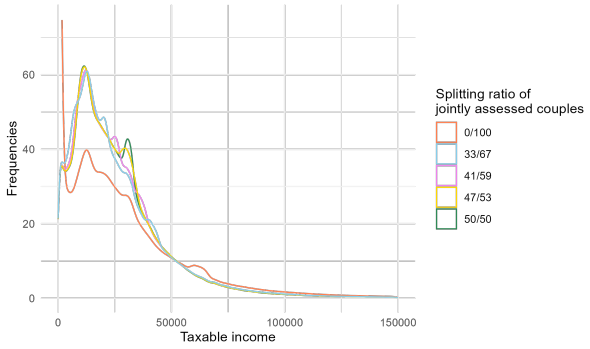
Figure A.1 shows the univariate distributions of taxable income under different splitting ratios. The unnaturally pronounced peak visible in the univariate distribution of taxable income for joint taxpayers (0/100 split) is due to jointly assessed tax payers. Dividing their taxable income according to an appropriate splitting ratio substantially reduces this effect, although the shape of the univariate distribution varies depending on the chosen ratio.

In our analysis, we focus on HITP, defined as the top 10% of the taxable income distribution. The threshold for being classified as HITP changes depending on the chosen splitting ratio. Table A.2 reports the corresponding threshold values.

Table A.2: 90th percentile of the income density depending on the splitting ratio

Split ratio	90% quantile
33/66	55,295 €
41/59	54,463 €
47/53	54,143 €
50/50	54,102 €

Figure A.1: Univariate income density for different splitting ratios



B Extended analysis of the density of income taxpayers and high-income taxpayers

As demonstrated in Sections 4.2 and 4.3, smoothed representations are preferable. To enable smooth regional distributions of ITPs and HITPs, kernel density estimates are used (see Section 3.1). Section B.2 will delve deeper into the analysis of the densities of these groups. First, the choice of bandwidth for kernel density estimation will be discussed in more detail. The effects of different calculation methods will be presented. Furthermore, in Section B.3 the effects of the splitting ratios (see Appendix A) in the context of determining HITPs will be analyzed. To facilitate comparisons between densities, Section B.1 introduces the Root Mean Integrated Squared Error as a

global measure and Jaccard similarity for densities as a local measure of comparison.

B.1 Comparison measures between two densities

The RMISE serves as a global comparison measure, computing the root of the mean squared deviation between two densities f_1 and f_2 across the area of interest. For this purpose, the area is discretized into $x_g,\ g = 1, \ldots,\ G$ grid points, where Δ_1 and Δ_2 represent the distance between the grid points along the coordinates. The RMISE is defined as

$$\text{RMISE} = \left(\int \big(f_1(x) - f_2(x)\big)^2 \, dx \right)^{1/2} \approx \sqrt{\sum_{x_g:\ g=1,\ldots,\ G} \big(f_1(x) - f_2(x)\big)^2 \Delta_1 \Delta_2} \ . \tag{10}$$

The Jaccard similarity for proportions (see equation (9)) behaves similarly to the Jaccard similarity for densities. To simplify the analysis, unlike the definition for proportions, a threshold is not directly specified; instead, the thresholds are defined based on quantile values of the given densities and hence, the definition of the sets U and V differ. To determine the similarity of the p% hotspots of densities f_1 and f_2, we consider the grid points x_g in U and V that correspond to the upper p% of the probability mass of f_1 and f_2, respectively. The Jaccard similarity for densities is then given by the equation (9).

B.2 Impact of bandwidth choice on the density of income taxpayers

The choice of the bandwidth matrix H is crucial for the quality of the kernel density estimation, see Section 3.1, as it significantly determines the bias-variance trade-off. The `ks` package in `R`, which was used in the implementation of the methods (see Sections 3.2 and 3.3), includes several methodologically different approaches to bandwidth determination, whose smoothing properties differ. Therefore, we will first introduce different bandwidth selection methods and then apply them on the data and measure its impact.

B.2.1 Methods for bandwidth selection

The *Plug-in method (PI)* minimizes an approximation of the integrated mean squared error (MISE), using a pilot density to estimate unknown terms such as $\int (\nabla^2 f(x))^2 dx$, where the Laplace Operator ∇^2 refers to the second partial derivative. This results in a bandwidth H_{PI}, which is often relatively small and leads to a detailed, less smoothed density estimate; for details see Wand et al. (1994).

The Normal-Scale (NS) method is based on the assumption that the underlying density is normally distributed and uses an analytical minimal solution of the asymptotic MISE (AMISE). The estimate known as Silverman's Rule of Thumb,

$$H_{\text{NS}} = n^{-\frac{1}{3}} \hat{\Sigma},$$

tends to provide the widest range compared to other approaches, and thus a highly smoothed estimate, where $\hat{\Sigma}$ represents the data-generated covariance matrix. For a detailed multivariate discussion, we refer to Chacón and Duong (2018), with the bivariate case illustrated here.

Least-squares cross-validation (LSCV) selects H by minimizing the integrated quadratic error via cross-validation:

$$\text{LSCV}(H) = \int \widehat{f}_H(x)^2 dx - \frac{2}{n} \sum_{i=1}^{n} \widehat{f}_{H,-i}(X_i),$$

where $\widehat{f}_{H,-i}$ denotes the leave-one-out estimation. This method tends to yield very small bandwidths. For more

information, see Gündüz and Karakoç (2023).

The *Smoothed-Cross-Validation (SCV)* method is a stabilization of LSCV, in which the first term is replaced by a smoothed version

$$\int \widehat{f}_{H'}(x)\widehat{f}_H(x)dx,$$

where H' is an additional smoothing bandwidth. SCV usually produces larger bandwidths than the LSCV method and thus a smoother estimate than LSCV yields; see Duong and Hazelton (2005) for details.

Nevertheless, the order of magnitude of the bandwidth estimate depends on the datasets.

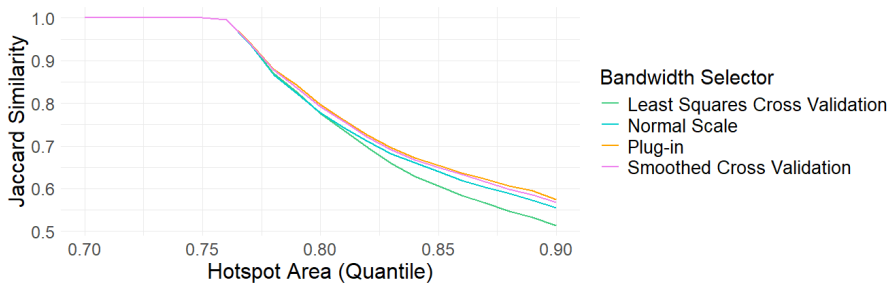
B.2.2 Effects of different bandwidth selectors on the density of the income taxpayers

The results of the kernel density estimation of the 10% sample of the ITPs over the Berlin city area are shown in Figure 4.1 for the bandwidth selected using the plug-in method. Other bandwidth selection methods lead to different estimates. The plug-in bandwidth for full data yields an estimate of approximately 170 meters along the x-axis (longitude) and approximately 150 meters along the y-axis (latitude). It should be noted that the data are based on the LAEA coordinate system (ETRS89, CRS3035), which provides a metric map projection system for Europe. The normal-scale bandwidth is considerably larger, whereas the bandwidths obtained via LSCV and SCV are of a similar magnitude, exhibiting a more pronounced rotation of the smoothing ellipse.

Compared to the unsmoothed counts, clearly recognizable deviations emerge, caused by the influence of the infrastructure, etc. We refer to Figures 4.1 and 4.2 for further details. These differences can be quantitatively represented using the Root Mean Integrated Squared Error (RMISE) and Jaccard similarity.

Figure B.1 shows the Jaccard similarities of the KDE estimates with different bandwidths relative to the frequencies. The horizontal axis indicates the p quantile range that defines the corresponding hotspot. The smaller the hotspot range under consideration, the greater the deviations between the smoothed estimate and the unsmoothed frequencies. Across all considered bandwidth methods, a similar pattern emerges.
In addition to local analyses, the global error measure shows similar deviations using different bandwidths. The RMISE between the frequency data and the kernel density estimates with plug-in bandwidth is $1.15 \cdot 10^{-3}$. The normal-scale bandwidth leads to a minimal increase in the RMISE of 0.2% and LSCV of 0.4%, while SCV shows a significant increase of approximately 12%. However, both LSCV and SCV are computationally intensive, limiting their practical use. Between the densities estimated with different bandwidth selection methods, the RMISE is on the order of 10^{-6}, which can be considered as very small. The effects of different bandwidths are marginal in our case, likely due to the high number of data points.

Figure B.1: Jaccard Similarity between unsmooth frequency data and smooth KDE estimates using different bandwidth selection methods



B.3 Impact of the splitting ratio on the density of the high income taxpayers

Analogous to the analysis of different bandwidths, the regional distributions are examined for varying splitting ratios of jointly assessed taxpayers. The exact taxable income for tax cases who are jointly assessed is unknown. Hence, the individual income of jointly assessed tax cases is approximated using a splitting ratio. However, this has little impact on the regional distribution of the HITP. The RMISE between the HITP frequency data and the kernel density estimate using the plug-in bandwidth for different splitting ratios is in the order of 10^{-3}.

Furthermore, local analyses using Jaccard similarity show that the hotspot similarity between the frquency data and the smoothed data is of a similar order of magnitude, regardless of the chosen splitting ratio (see Figure B.2). Due to the influence of grid cell size and infrastructure, the similarity of the hotspots decreases in regions with high densities or frequencies.

Figure B.3 shows the similarity between hotspots across different splitting ratios using KDE as a basis. This similarity is consistently above 95%, suggesting very similar kernel density estimates despite the different splitting ratios. Nevertheless, it is evident that split ratios that differ more significantly in percentage terms, e.g., 50/50 versus 33/67 split, show the least agreement.

Figure B.2: Jaccard Similarity between unsmooth frequency data of HITP and smooth KDE estimates using different splitting ratios

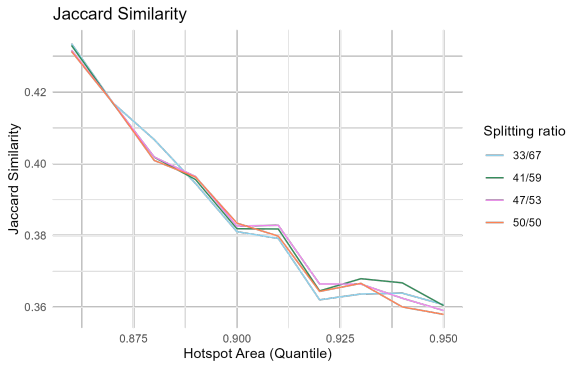
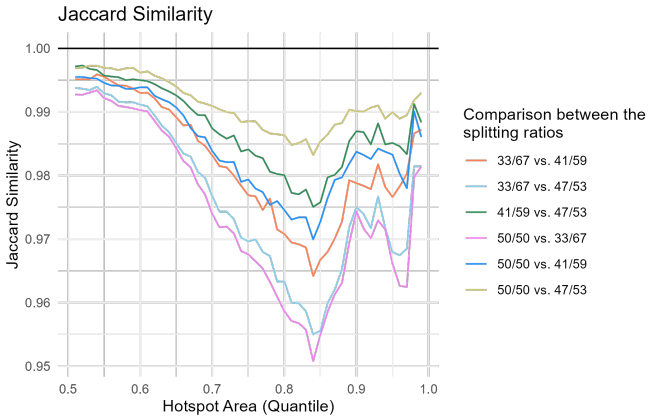


Figure B.3: Jaccard Similarity between smoothed KDEs across different splitting ratios



C Aggregation of income taxpayers

The preliminary step of disaggregating proportion values consists of systematically distributing an aggregated population across the underlying area. Here, the aim is to represent a smooth density on anonymized ITPs and quantifying the error caused by anonymization.

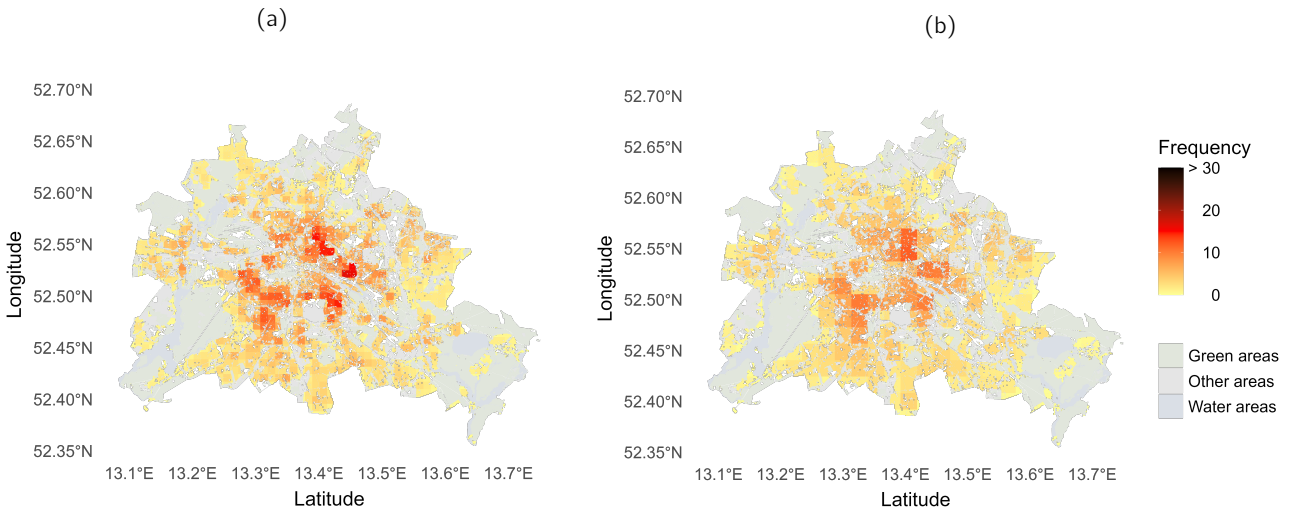
The proposed measurement error model from Section 3.2 aims to obtain smooth estimates through kernel density estimations, assuming that the measurement error process is known. For this, aggregated data are observed, and knowledge of the aggregation process as a measurement error is required, i.e., the centers of the regions or the rectangle, and the shape of the region as a polygon P_a, for $a = 1, \ldots, A$. For details on the numerical implementation, see Groß et al. (2016).

C.1 Application of the measurement error model to aggregated income taxpayers

In the disaggregation of proportion values, the measurement error model is first applied to the entire population, after which a subpopulation with the relevant characteristic is selected. In this section, we focus exclusively on the disaggregation of the ITPs. As it was the case with the proportions, in addition to the goal of achieving an

improved representation, it is also crucial to examine whether the associated information loss can be reduced.

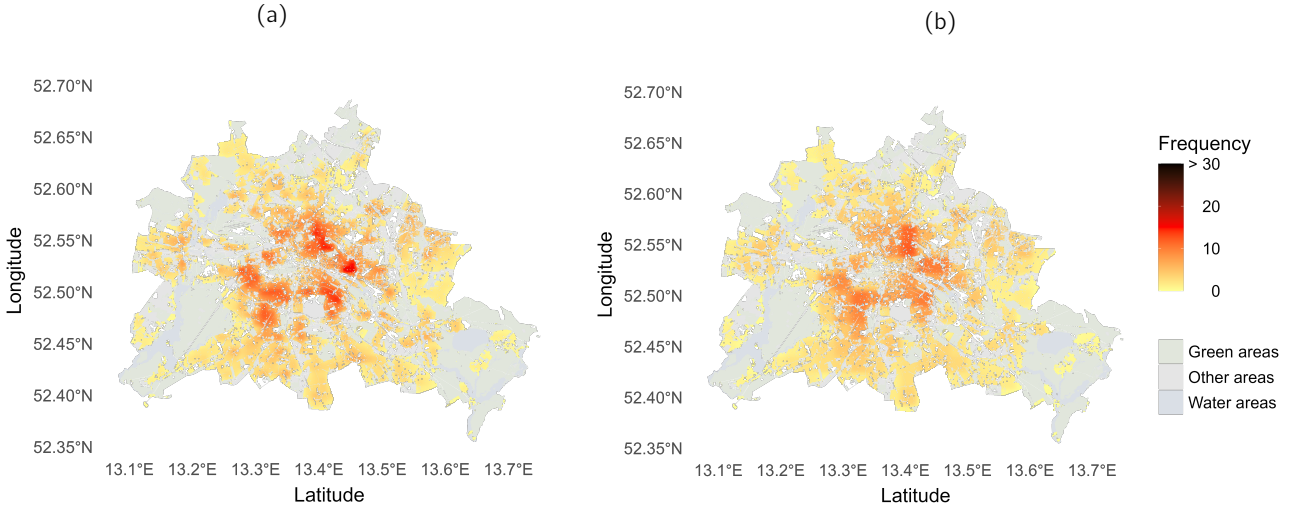
Figure C.1: Illustration of the aggregated ITP for an (a) 800 × 800 and (b) 1800 × 1800 meter aggregation rectangles



Since the residential location information for roughly two million ITPs in Berlin is available at the address level, aggregation is performed by rounding the geocoordinates to a predefined rectangular grid. This allows for very fine increments in the size of the rectangles, thus systematically illustrating the information loss during aggregation. The edge length of the rectangles varies between 200 and 2000 meters in 100-meter increments. Figure C.1 shows the aggregated data on rectangles with edge lengths of 800 and 1800 meters.

Rounding the geoccordinates of ITPs, as commonly visualized in choropleth maps, leads to discontinuities at the edges of the aggregation units. For small aggregation units, it is at least possible to discern trends regarding areas with a high concentration of ITPs. However, the abrupt changes in values appear irregular and visually uninformative. For larger aggregation units, the resulting information loss becomes clearly apparent. Notable, the values in the legend of Figure C.1 correspond again to 100 × 100 meter grids and uninhabited areas are not depicted, making some aggregation units harder to detect.

Figure C.2: Illustration of the smoothed density using the measurement error model, which used the number of ITP aggregated on an (a) 800 × 800 and (b) 1800 × 1800 meter aggregation rectangles



The measurement error model under aggregation uses information about the position of the centers and the shape of the aggregation unit as well as the associated population value of the ITPs. A smoothed estimate is achieved by iteratively drawing a pseudo-sample from the aggregates to generate spatially dispersed coordinates representing the ITP residences. Taking the measurement error into account, an estimated ITP density is obtained. Figure C.2 shows the estimates of the smoothed densities from aggregated information for the aggregates with edge lengths of 800 and 1800 meters.

It is clearly evident that the algorithm generates smoothed densities of the ITP. However, comparing the KDE based on the original coordinates (see Figure 4.1b) with the densities based on aggregated information reveals that while large-scale local clusters could be recognized, detailed structures were lost. When aggregation was performed over small-scale clusters, these cannot be recovered.

Figure C.3: RMISE comparing the smooth density of ITP with the aggregated density (red) and smooth density from aggregated data using the measurement error model over Berlin across varying aggregation rectangle sizes

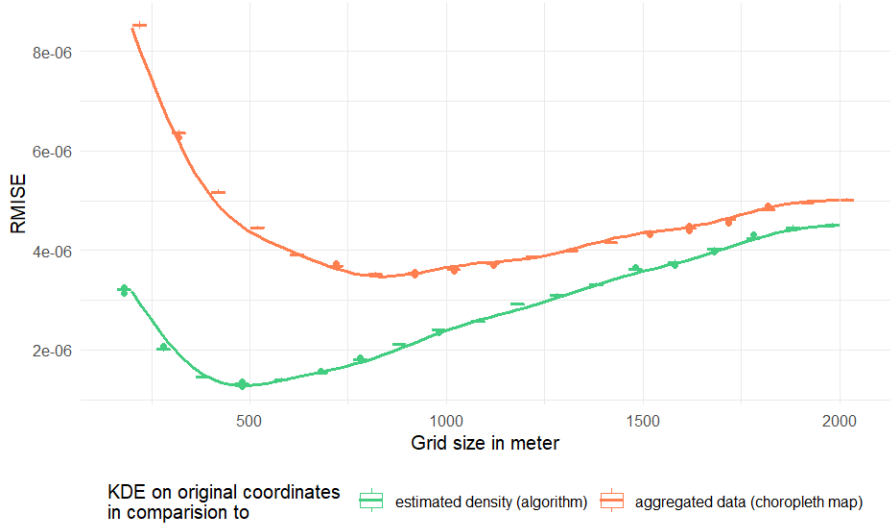


Figure C.3 shows the RMISE, which is used to measure error for the cases considered. In all scenarios, 100 simulations were performed on a 10% subsample. The subsample is highly representative, which explains why the range of the displayed boxplots is very small and thus difficult to discern. As a reference, the KDE of the exact data is used, providing a smooth representation of the ITP distribution. The orange line depicts the comparison with the aggregated data, where the same aggregation value applies to the entire aggregation unit. The green line illustrates the comparison with the smoothed estimates obtained from the measurement error model, which uses the aggregated data as its basis. Across all cases, the algorithm produces an improved representation relative to the aggregated data. Notably, the error curve exhibits a distinct "J"-shaped pattern. It is worth noting that only when the aggregation units reach a rectangular size of 500 m per edge are there sufficient aggregated evaluation points to make the use of the algorithm meaningful.

References

Baddeley, Adrian, Ege Rubak, and Rolf Turner (2015). *Spatial Point Patterns: Methodology and Applications with R*. Milton: CRC Press LLC.

Behnisch, Martin et al. (June 2013). "Using quadtree representations in building stock visualization and analysis". In: *Erdkunde* 67.2, pp. 151–166.

Bensmann, Felix et al. (July 2020). "An infrastructure for spatial linking of survey data". en. In: *Data Science Journal* 19.

Brenzel, Hanna and Kathrin Gebers (2020). "Werkstattbericht: Georeferenzierung im Statistischen Verbund". In: *WISTA – Wirtschaft und Statistik* 72.6, pp. 48–57. ISSN: 1619-2907.

Burian, Jaroslav, Jan Zapletal, and Vít Pászto (June 2022). "Disaggregator – a tool for the aggregation and disaggregation of spatial data". en. In: *Earth Science Informatics* 15.2, pp. 1323–1339.

Chacón, José E and Tarn Duong (2018). *Multivariate kernel smoothing and its applications*. Chapman and Hall/CRC.

Diggle, Peter (1985). "A kernel method for smoothing point process data". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 34.2, p. 138.

Duong, Tarn and Martin L Hazelton (2005). "Cross-validation bandwidth matrices for multivariate kernel density estimation". In: *Scandinavian Journal of Statistics* 32.3, pp. 485–506.

Eurostat (2025). *Statistical Confidentiality*. `https://ec.europa.eu/eurostat/de/about-us/statistical-confidentiality`. Accessed: 2026-01-22.

Federal Agency for Cartography and Geodesy in Germany (2019). *Dataset of geographical grids for Germany in ETRS89-LAEA*. Accessed: 2026-01-22. URL: `https://mis.bkg.bund.de/trefferanzeige?docuuid=02A7E63D-CAAA-4DED-B6FF-1F1E73FAF883`.

Fraser, Bruce and Janice Wooten (2005). *A Proposed Method for Confidentialising Tabular Output to Protect Against Differencing*. Working paper, Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Geneva, Switzerland, 9–11 November 2005. Accessed: 2026-01-22. URL: `https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2005/wp.35.e.pdf`.

Gril, Lorena et al. (2025). "Kernel Heaping – Kernel Density Estimation from Regional Aggregates via Measurement Error Model". In: *The R Journal* 16.3, pp. 115–133.

Groß, Marcus and Ulrich Rendtel (Sept. 2016). "Kernel Density Estimation for Heaped Data". In: *Journal of Survey Statistics and Methodology* 4.3, pp. 339–361.

Groß, Marcus et al. (Feb. 2016). "Estimating the Density of Ethnic Minorities and Aged People in Berlin: Multivariate Kernel Density Estimation Applied to Sensitive Georeferenced Administrative Data Protected Via Measurement Error". In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 180.1, pp. 161–183.

Gündüz, Necla and Şule Karakoç (2023). "Optimal bandwidth selection methods with application to wind speed distribution". In: *Mathematics* 11.21, p. 4478.

Hall, Peter and Berwin A Turlach (Mar. 1999). "Reducing bias in curve estimation by use of weights". en. In: *Computational Statistics & Data Analysis* 30.1, pp. 67–86.

Härdle, Wolfgang (1990). *Applied nonparametric regression*. Cambridge: Cambridge University Press.

Hundepool, Anco et al. (2018). *Handbook on Statistical Disclosure Control*. Accessed: 2026-01-22. SDC Tools Project. URL: `http://sdctools.github.io/HandbookSDC/Handbook-on-Statistical-Disclosure-Control.pdf`.

Izenman, Alan Julian (1991). "Recent Developments in Nonparametric Density Estimation". In: *Journal of the American Statistical Association* 86.413, pp. 205–224.

Jones, Chris M. (Sept. 1993). "Simple boundary correction for kernel density estimation". In: *Statistics and Computing* 3.3, pp. 135–146.

Li, Songnian et al. (2016). "Geospatial big data handling theory and methods: A review and research challenges". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 115, pp. 119–133.

Ricciato, Fabio and Angelo Coluccia (June 2023). "On the estimation of spatial density from mobile network operator data". In: *IEEE Transactions on Mobile Computing* 22.6, pp. 3541–3557.

Rushton, Gerard et al. (Feb. 2006). "Geocoding in cancer research: a review". en. In: *American Journal of Preventive Medicine* 30.2 Suppl, S16–24.

Schweers, Stefan et al. (2016). "Conceptualizing a Spatial Data Infrastructure for the Social Sciences: An Example from Germany". In: *Journal of Map & Geography Libraries* 12.1, pp. 100–126.

Scott, David W. and Simon J. Sheather (1985). "Kernel density estimation with binned data". In: *Communications in Statistics – Theory and Methods* 14.6, pp. 1353–1359.

Senatsverwaltung Berlin (2026). *Daten Berlin*. `http://daten.berlin.de`. Accessed: 2026-01-22. URL: `http://daten.berlin.de`.

Silverman, Bernard W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman & Hall.

Statistisches Bundesamt (Destatis) (2021). *Lohn- und Einkommensteuerstatistik*. Accessed: 2026-01-21. URL: `http://destatis.de/DE/Methoden/Qualitaet/Qualitaetsberichte/Steuern/lohn-und-einkommensteuer.pdf?__blob=publicationFile&v=12`.

— (2024). *Zensus-Atlas 2022: Interaktive kartografische Anwendung auf Gitterzellenbasis*. `https://atlas.zensus2022.de/`. Accessed: 2026-01-22. Statistische Ämter von Bund und Ländern.

— (2025a). *Lohn- und Einkommensteuer*. `https://www.destatis.de/DE/Themen/Staat/Steuern/Lohnsteuer-Einkommensteuer/_inhalt.html`. Accessed: 2026-01-21. URL: `https://www.destatis.de/DE/Themen/Staat/Steuern/Lohnsteuer-Einkommensteuer/_inhalt.html`.

— (2025b). *Unadjusted Gender Pay Gap by Länder from 2014 to 2025*. `https://www.destatis.de/EN/Themes/Labour/Earnings/GenderPayGap/Tables/ugpg-02-by-laender-at2014.html`. Accessed: 2026-01-21. URL: `https://www.destatis.de/EN/Themes/Labour/Earnings/GenderPayGap/Tables/ugpg-02-by-laender-at2014.html`.

Strobl, Josef (2004). "Hierarchische Aggregation: Detailinformation versus Datenschutz am Beispiel adressbezogen georeferenzierter Datensätze". In: *Salzburger Geographische Arbeiten*, pp. 163–171.

Wand, Matt P. and Chris Jones (1994). "Multivariate plug-in bandwidth selection." In: *Computational Statistics* 9.2, pp. 97–116.

Wilson, Benjamin, Neal Wilson, and Sierra Martin (Oct. 2021). "Using GIS to advance social economics research: Geocoding, aggregation, and spatial thinking". en. In: *Forum for Social Economics* 50.4, pp. 480–504.

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

2026 erschienen:

2026/1 Hundsdoerfer, Jochen und Maren Löwe: How Do Value Added Taxes Affest Wages and Labor?
FACTS