Estimating the density of ethnic minorities and aged people in Berlin: Multivariate kernel density estimation applied to sensitive geo-referenced administrative data protected via measurement error

Marcus Groß [*] Ulrich Rendtel [*] Timo Schmid [*]
Sebastian Schmon[†] Nikos Tzavidis[‡]

Abstract

Modern systems of official statistics require the timely estimation of area-specific densities of sub-populations. Ideally estimates should be based on precise geo-coded information, which is not available due to confidentiality constraints. One approach for ensuring confidentiality is by rounding the geo-coordinates. We propose multivariate non-parametric kernel density estimation that reverses the rounding process by using a measurement error model. The methodology is applied to the Berlin register of residents for deriving density estimates of ethnic minorities and aged people. Estimates are used for identifying areas with a need for new advisory centres for migrants and infrastructure for older people.

Keywords: Ageing; Binned data; Ethnic segregation; Non-parametric estimation; Official statistics.

1 Introduction

Modern systems of official statistics require the estimation of area-specific densities of sub-populations. In large cities researchers may be interested in identifying areas with high density of ethnic minorities or areas with high density of aged people. The focus can be even more specific for example, on density estimates of school age children of ethnic minority background. In this paper the term ethnic minority will be used to define the part of the population with migration background. Estimates of this type can be used by researchers in Government Departments and other organisations for designing and implementing targeted policies.

[*] Institute for Statistics and Econometrics, Freie Universität Berlin, Germany, marcus.gross@fu-berlin.de, ulrich.rendtel@fu-berlin.de, timo.schmid@fu-berlin.de

[†] Department of Statistics, University of Oxford, UK, schmon@stats.ox.ac.uk

[‡] Southampton Statistical Sciences Research Institute, University of Southampton, UK, n.tzavidis@soton.ac.uk

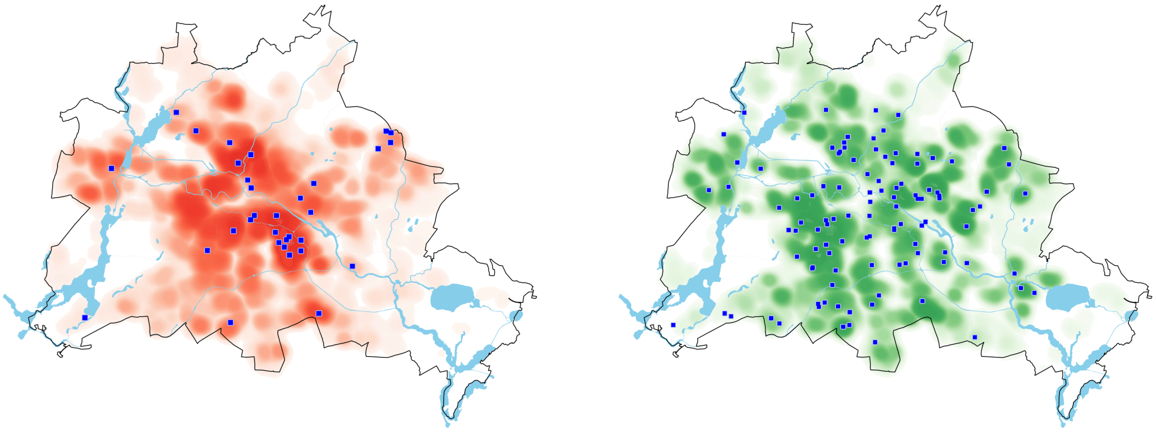

Figure 1: Density estimates of the population of ethnic minority background (left map) and of the population aged 60 or above in Berlin (right map). The blue points (left map) show the spatial distribution of advisory centres for migrants. The blue points (right map) show the spatial distribution of care homes.

To motivate the methodology we propose in this paper, we start by presenting two maps in Figure 1. The left map presents an estimate of the density of the population of ethnic minority background in Berlin. The right map presents an estimate of the density of the population aged 60 or over in Berlin. The blue points superimposed on the left map show the spatial distribution of advisory centres in Berlin. These are centres that provide assistance for migrants in Berlin. The blue points superimposed on the right map show the spatial distribution of care homes in Berlin. Both kernel density estimation plots in Figure 1 have been produced by using real data from the Berlin register, which is a register of residents in all Berlin household addresses that contains exact geo-coded coordinates. At this point we must mention that the register data is available only to the data host in a safe environment. Hence, for producing Figure 1 we had to rely on collaborating with staff at the Berlin-Brandenburg Statistics Office who monitored the in-house use of the data. Maps such as those we presented in Figure 1 can be very useful for planning purposes. For example, city councils can use the density estimation plots to decide where new advisory centres for migrants are mostly needed or for deciding in which areas to offer planning permissions for opening new care homes. Register databases are updated on a frequent basis and hence their timeliness is better than that of alternative sources of data for example, Census data.

The statistical problem we face in this paper is created by the fact that the register with the exact coordinates used for producing the maps in Figure 1 is not publicly available. Access to such data is impeded by confidentiality constraints (VanWey et al., 2005) and this holds true also for the Berlin register data. It is easy to see why confidentiality constraints are in place. The availability of precise geo-coding alongside information on demographic characteristics can increase the disclosure risk in particular for sensitive sub-groups of the population such as ethnic minorities. Restricted access to sensitive data may not only apply to users working outside the data host but also to researchers working for the data host or for related organisations for example, Government Departments. As we will see in this paper, in the case of the Berlin register data specific procedures are used to ensure confidentiality of the sensitive data. Nevertheless, policies that govern access to sensitive data are country-specific. Other countries that have a long tradition of maintaining register geo-coded data are the Scandinavian ones for example, Norway and Finland. However, access to and use of such data is restricted and these restrictions are decided by the data host in each country.

The host of the data can offer access, possibly in a safe setting, to geo-coded data whilst ensuring confidentiality. One way to achieve this is by introducing measurement error to longitudes and latitudes (Armstrong et al. 1999; Ozonoff et al. 2007 or Rushton et al. 2007). However, this raises the following question. Can we derive precise density estimates of the sub-groups of interest by using data that has been subjected to disclosure control via the introduction of measurement error in the geographic coordinates? The present paper proposes non-parametric multivariate density estimation in the presence of measurement error in the geographic coordinates. The aim is to investigate how the precision of density estimates produced by using coarsened data and the use of a non-parametric statistical methodology for reversing the measurement error process compares to density estimates produced by using the exact geo-referenced data. At this point we should make clear that the paper does not discuss whether the released geo-referenced information makes identification possible. Instead, we assume that the parameters of the disclosure control process are decided by the data provider. For a discussion on the effectiveness of anonymisation techniques, we refer the reader to Kwan et al. (2004).

Scott and Sheather (1985) used $Naive$ density estimation methods that disregard the presence of rounding. To account for rounding Härdle and Scott (1992) introduced a kernel-type estimator based on weighted averages of rounded data points and Minnotte (1998) developed an approach of histogram smoothing. An iterative estimation scheme presented by Blower and Kelsall (2002) ensures non-negative estimates and can potentially be applied to multivariate data as well. A recent publication of Xu (2014) extends this approach to asymmetric kernels. However, the bandwidth selection which is crucial in kernel density estimation is done with a rather ad-hoc approach on the binned data. Wang and Wertelecki (2013) proposed a parametric and a non-parametric kernel density estimator for rounded data but considered only the univariate case. Wang and Wertelecki (2013) showed that using a $Naive$ kernel density estimator to rounded data with standard bandwidth selection may lead to poor results for large rounding intervals and large sample sizes.

An alternative idea, explored in this paper, is to interpret rounding as a measurement error process and to formulate the problem by using measurement error models (Carroll et al., 2010; Fuller, 2009). For classical additive error models the problem can be regarded as density deconvolution and can be solved using Fourier methods (Stefanski and Carroll, 1990; Zhang, 1990). The topic of density deconvolution has been extensively studied and literature has focused on optimal convergence rates (Fan et al., 1991), different error distributions such as Gaussian or uniform distributions (Feuerverger et al., 2008) and choice of an optimal bandwidth (Delaigle and Gijbels, 2004). Moreover, the case of additive Berkson errors (Berkson, 1950) in the context of non-parametric density estimation has been investigated. Delaigle (2007, 2014) proposed a density estimator which does not require any bandwidth choice and converges at a parametric rate but with the drawback of producing spiky estimates with high variance when the measurement error is rather low. A recent paper by Long et al. (2014) empirically compares the estimator of Delaigle (2007, 2014) to two novel approaches for multivariate kernel density estimation contaminated with Gaussian Berkson error and states that one of them shows superior performance. However, rounding error can neither be classified as classical nor Berkson additive error structure as the error is neither independent of the true coordinate nor the rounded one. Nevertheless, a Berkson model with uniform error distribution can be used as an approximation (Wang and Wertelecki, 2013). In this case the estimator by Delaigle (2007) is a bivariate histogram type estimator. When the rounding error, which governs the binwidth, is high the estimator proposed by Delaigle (2007) can be biased. Therefore, in this paper we develop a method that correctly specifies the measurement error model under rounding.

From a methodological perspective the present article proposes a novel approach to multivariate non-parametric kernel density estimation in the presence of rounding errors used to ensure data confidentiality. The main advantage of the proposed methodology, compared to alternative methodologies, is that under our approach the bandwidth is derived as part of the estimation process. Moreover, our method is very easy to implement and works regardless of the dimension, the kernel and the bandwidth selection method.

In this paper we assume only the availability of register geo-coded data with measurement error in the geographic coordinates. Hence, conventional estimation methods that combine Census/register data with survey data are not applicable in this case. In this paper we use the Berlin register data, a complete enumeration of the entire Berlin population in private households, for illustrating how to derive precise density estimates of sensitive groups in the presence of measurement error in two applications.

The first application aims at estimating the density of the Berlin population that is of ethnic minority background. The focus on this application is motivated by the debate on integration/segregation of migrants. Residential segregation describes the phenomenon of a separation of residents according to certain characteristics such as ethnicity. Recent literature suggests that higher levels of segregation are linked with higher crime rates and lower health and educational outcomes (Peterson et al., 2008; Card and Rothstein, 2007; Acevedo-Garcia et al., 2003). To prevent the segregation of ethnic minorities it is necessary to assist these groups with integration programmes offered by advisory centres. Programmes of this kind should be established in areas with high density of ethnic minorities. For the purposes of this application we study the current location of advisory centres in relation to density estimates and identify areas where more support is potentially needed.

The second application relates to the provision of social services for the elderly and urban planing in the context of changing demographics. Longer life expectancy and declining birth rates lead to an ageing population, which needs to be accounted for in urban and social planning. For example, the German National Statistical Institute (Destatis, 2009) predicts the ratio of people over 65 to rise from 20% in 2008 to 34% in 2060. This is a common issue for other industrialised countries too. To ensure the wellbeing of the elderly and to secure adequate and affordable support for this group it is necessary to analyze where the elderly live. Gorr et al. (2001) used the density of the elderly population as a basis for a spatial decision support system for home-delivered services (meals on wheels). Further challenges arise in urban planing, where an ageing population requires easy access to buildings, public services and public transportation. Shortcomings in urban development can be analyzed by comparing the density of the elderly population against those characteristics (Verma, 2014). In addition, many elderly people decide to live in a retirement home. To secure adequate and affordable support for the elderly population it is necessary to establish services where needed. The methodology we propose in this paper is also used for providing precise density estimates of the elderly population in the Berlin area. For both applications the sensitivity of density estimation to the severity of the rounding error process is studied and the proposed methodology is contrasted to a $Naive$ kernel density estimator which disregards the presence of measurement error.

The structure of the paper is as follows. In Section 2 we describe the Berlin register data. In Section 3 we review multivariate kernel density estimation in the presence of measurement error. A multivariate kernel density estimator is proposed and the computational details of the proposed method are described. In Section 4 we present the results of the two applications by using the Berlin register data. In Section 5 we empirically evaluate the performance of the proposed methodology under different assumptions for the rounding error process with data generated from known bivariate densities. The precision of the density estimates provided by the proposed methodology is contrasted to the precision of the estimates derived by (a) using a $Naive$ kernel density estimator that disregards the presence of rounding error and (b) alternative approaches that have been proposed in the literature. Finally, in Section 6 we conclude the paper with some final remarks.

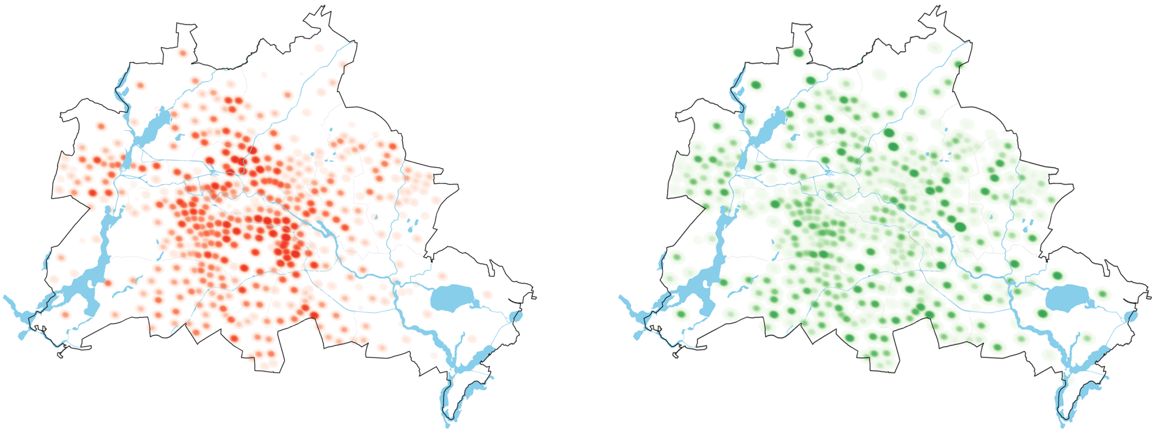

Figure 2: Density estimates of the population with ethnic minority background in Berlin (left map) and of the population aged 60 or above in Berlin (right map) based on the publicly available data.

2 The Berlin register data

The statistical problem we face in this paper is motivated by the Berlin register of residents dataset, which comprises all Berlin household addresses and contains exact geo-coded coordinates. Such a comprehensive data set is gathered because of German legislation. In particular, registration at the local residents' office is compulsory in Germany and is carried out by the federal state authorities. In the federal city state of Berlin registration is regulated by the Berlin registration law. This law requires every person who moves into a new residential unit in Berlin to be registered in person within one week.

This register is not publicly available because of the detailed geo-coded information it contains. However, a version of the register data is publicly available as part of the Open Data initiative in Berlin (`http://daten.berlin.de`), an initiative that aims at using data for improving urban development. The open dataset includes aggregates for the 447 lowest urban planning areas, the so-called LORs ('Lebensweltlich orientierte Räume'), with coordinates given by the centroid of these areas. This is a discrete and possibly arbitrary demarcation. The discreteness of the demarcation is apparent in Figure 2, which shows kernel density estimates of the population of ethnic minorities (left map) and of the population aged 60 or over (right map) in Berlin by using the publicly available data. A main aim of the present paper is to derive precise density estimates of population groups by using a more flexible definition of geographic demarcation. This in turn may provide more useful information to local authorities than the currently available LOR demarcation.

An alternative to the currently available data, and one explored by the data host, is to generate a grid-based version of the data that is independent from the somewhat arbitrary geometry of the LORs. In this case the grid-aggregates can be interpreted as the result of rounding geo-coded data for ensuring data confidentiality. Here each point of the grid defines a square-shaped area around the grid point with a longitude and latitude increment equal to the grid length. Then the value of the variable of interest is the aggregate of the values with exact geo-coordinates over the area surrounding the grid point. In fact, the LOR demarcation in Berlin can be thought of as the process of rounding the geo-referenced data by using grids of average size 2000 meters by 2000 meters. The methodology we propose in this paper attempts to reverse the rounding process for deriving estimates that are more precise than density estimates that ignore the measurement error process and relate to a more flexible definition of geographic demarcation.

The data that we have access to in this paper contains all 308,754 Berlin household addresses on the 31st of December 2012 with the exact geo-coded coordinates subject to different degrees of rounding error. One of the scenarios we explore is rounding by using grids of size 2000 meters by 2000 meters that approximately correspond to the LOR demarcation. The location is measured by (Soldner)-coordinates in meters. The original (without rounding error) data includes the total number of residents (Berlin Total) at their principal residence and the number of persons according to some key demographic characteristics. The first demographic variable is the migration background (Migration) of individuals defined by the number of people that are of (a) non-German nationality, (b) German nationality but born abroad and (c) non-German nationality who changed their nationality into German at the coordinates of the principal household address. The definition of this variable is further refined by the number of individuals of migration background from Turkey (Migration Turkey) or Vietnam (Migration Vietnam). The second demographic variable is age (age over 60) defined by the number of individuals who are older than 60 years old. The density estimates of the subgroups of interest that are produced by using the proposed methodology are contrasted to maps of the corresponding densities produced by using the data with the exact geo-coded coordinates. The use of these maps has been approved by the data host, the Berlin-Brandenburg Statistics Office.

Table 1 presents summary statistics of the number of residents living at a household address of the key variables based on the exact geo-coded data. Due to confidentiality restrictions we are not allowed to publish the maximum number of residents living at a household address. The average of individuals living at a household address in Berlin is 11.24 leading to a total population of 3,469,619 (registered) inhabitants. Note that a household address in the data is defined for example, as an entire block of apartments. Around 27% of the total population are of migration background and around 24.8% of the population are older than 60 years. The average number of residents of migration background is 3.07 with a median of 0, whereas the average number of individuals above 60 years of age is 2.78 with a median of 1. This gives a first indication that inhabitants with migration background are more clustered compared to older people in Berlin.

Table 1: Summary statistics of the number of residents living at a household address.

	Sum	Min.	1st Qu.	Median	Mean	3rd Qu.
Berlin Total	3,469,619	1	2	4	11.24	15
Migration	949,184	0	0	0	3.07	3
Migration Vietnam	21,637	0	0	0	0.07	0
Migration Turkey	176,738	0	0	0	0.57	0
Age over 60	859,170	0	0	1	2.78	3

3 Multivariate kernel density estimation in the presence of measurement error

In this section we propose an approach to non-parametric multivariate density estimation in the presence of measurement error in particular, rounding of the geographical coordinates used for disclosure control of sensitive data. Multivariate kernel density estimation is introduced in Section 3.1. In Section 3.2 we investigate kernel density estimation in the presence of measurement error and in Section 3.3 we present a model that corrects for measurement error in multivariate kernel density estimation. Estimation and the computational details of the algorithm we use for implementing the proposed model are described in Section 3.4.

3.1 Multivariate kernel density estimation

Kernel density estimation as a non-parametric approach is an important tool in exploratory data analysis. Multivariate kernel density estimation attempts to estimate the joint probability distribution for two or more continuous variables. This method has the advantage of producing smooth density estimates compared to a histogram whose appearance heavily depends on the bin's breakpoints. Let $X = \{X_1, X_2, \dots, X_n\}$ denote a sample of size n from a multivariate random variable with unknown density $f(x)$. In the following, we only consider the two-dimensional case without loss of generality such that $x = (x_1, x_2)$. Thus, X_i, $i = 1, \dots, n$ is given by (X_{i1}, X_{i2}), where – in our application – X_{i1} and X_{i2} denote longitude- and latitude- coordinates, respectively.

The multivariate kernel density estimator at point x is given by

$$\hat{f}_H(x) = \frac{1}{n|H|^{\frac{1}{2}}} \sum_{i=1}^{n} K\left(H^{-\frac{1}{2}}(x - X_i)\right), \tag{1}$$

where $K(\cdot)$ is a multivariate kernel function, H denotes a symmetric positive definite bandwidth matrix and $|\cdot|$ denotes the determinant. A standard choice for $K(\cdot)$, used throughout this paper, is the multivariate Gaussian kernel. The choice of bandwidth H is crucial for the performance of a kernel density estimator. Approaches for bandwidth selection have been widely discussed in the literature. A popular strategy is to choose H by minimizing the asymptotic mean integrated squared error (AMISE) through plug-in or cross-validation methods (Izenman, 1991 or Silverman, 1986). In the univariate case we refer the reader to Marron (1987) or Jones et al. (1996). Wand and Jones (1994) discussed the choice of the bandwidth in the multivariate case by using a plug-in estimator. The approach by Wand and Jones (1994) is the one we use for bandwidth selection in this paper.

3.2 Rounding and kernel density estimation

By introducing rounding for achieving anonymisation of sensitive data the true values $X = \{X_1, X_2, \ldots, X_n\}$, the exact geographical coordinates, are lost. Instead, only the rounded (contaminated by measurement error) values, denoted by $W = \{W_1, W_2, \ldots, W_n\}$, are available. As a consequence the data is concentrated on a grid of points. Using a $Naive$ kernel density estimator that ignores the rounding process by replacing the true values X_i by the rounded values W_i in (1) may lead to a spiky density that is not close to the density of the uncontaminated (true) data. This effect becomes more pronounced with increasing sample size. In particular, as the bandwidth determinant $|H|$ is decreasing with higher sample size this causes higher density estimates on the grid points and lower in between the grid points.

The process of rounding means that the true, unknown, values $X_i = (X_{i1}, X_{i2})$ given the rounded values $W_i = (W_{i1}, W_{i2})$ are distributed in a rectangle with W_i in its center,

$$\left[W_{i1} - \frac{1}{2}r, W_{i1} + \frac{1}{2}r\right] \times \left[W_{i2} - \frac{1}{2}r, W_{i2} + \frac{1}{2}r\right]. \tag{2}$$

The value r denotes the rounding parameter. For instance, the data is rounded to the next integer for $r = 1$.

3.3 The Model

A model for the density $f(x)$ could be formulated parametrically, for example by a multivariate Gaussian distribution, or non-parametrically either by a mixture of parametric distributions (Escobar and West, 1995; Gelfand et al., 2005) or by using multivariate kernel density estimation as introduced in Section 3.1. As discussed in Section 3.2, the true values X_i are lost because of the rounding process and only the rounded values W_i are observed. However, we still aim to estimate the density $f(x)$ – from which our sample X is drawn – only by using the rounded values W_i. Under the assumption that the rounding/anonymisation process of the X_i is known, we are able to formulate a measurement error model $\pi(W|X)$ for W. In particular, the measurement error model $\pi(W|X)$ for rounding is defined by a product of Dirac distributions, $\pi(W|X) = \prod_{i=1}^{n} \pi(W_i|X_i)$, with

$$\pi(W_i|X_i) = \begin{cases} 1 & \text{for } X_i \in [W_{i1} - \frac{1}{2}r, W_{i1} + \frac{1}{2}r] \times [W_{i2} - \frac{1}{2}r, W_{i2} + \frac{1}{2}r] \\ 0 & \text{else.} \end{cases} \tag{3}$$

From the Bayes theorem it follows that $\pi(X|W) \propto \pi(W|X)\pi(X)$. Utilizing this formulation we can draw pseudo samples (imputations) of the X_i from $\pi(X_i|W_i)$ which enables us to estimate $f(x)$. As $\pi(X) = \prod_{i=1}^{n} f(X_i)$ is initially unknown we propose an iterative procedure, which uses an initial estimate of $f(x)$ based on the W_i followed by alternating simulations of X from $\pi(X|W)$ and re-estimation of $\pi(X)$ until convergence. The following subsection gives further details about the exact implementation of the algorithm and discusses how this can be viewed as a variant of the Expectation-Maximization (EM) algorithm (Dempster et al., 1977).

3.4 Estimation and Computational details

As discussed in the previous subsection, for fitting the model we need to draw pseudo samples of the X_i. The conditional distribution of the X_i given the rounded values W_i is the following:

$$\pi(X_i|W_i) \propto I(W_{i1}-\frac{1}{2}r \leq X_{i1} \leq W_{i1}+\frac{1}{2}r) \times I(W_{i2}-\frac{1}{2}r \leq X_{i2} \leq W_{i2}+\frac{1}{2}r) \times f(X_i), \quad (4)$$

where $I(\cdot)$ denotes the indicator function. The conditional distribution of X_i is the product of a uniform distribution on the square with side length r around W_i and density $f(x)$. As the density $f(x)$ is unknown it is replaced by an estimate, which is the multivariate kernel density estimator $\hat{f}_H(x)$ defined in (1). In particular, X_i is repeatedly drawn from the square of side length r around W_i using the current density estimate $\hat{f}_H(x)$ as a sampling weight. The steps of the algorithm are described below.

1. Get a pilot estimate of $f(x)$ by setting H to $\begin{pmatrix} l & 0 \\ 0 & l \end{pmatrix}$, where l is a sufficiently *large* value such that no rounding spikes occur.

2. Evaluate the density estimate $\hat{f}_H(x)$ on an equally-spaced fine grid $G = z_1 \times z_2$ (with $G = \{g_1, \ldots, g_m\}$, gridwidth δ_g and
$z_1 = \left\{\min_i(W_{i1}) - \frac{1}{2}r, \min_i(W_{i1}) - \frac{1}{2}r + \delta_g, \ldots, \max_i(W_{i1}) + \frac{1}{2}r\right\}$,
$z_2 = \left\{\min_i(W_{i2}) - \frac{1}{2}r, \min_i(W_{i2}) - \frac{1}{2}r + \delta_g, \ldots, \max_i(W_{i2}) + \frac{1}{2}r\right\}$ $(i = 1, \ldots, n)$),
where r denotes the rounding parameter introduced in Section 3.2.

3. Sample from $\pi(X_i|W_i)$ by drawing a sample $X_i^S = \left(X_{1i}^S, X_{2i}^S\right)$ randomly from $\left(z_1 \in [W_{i1} - \frac{1}{2}r, W_{i1} + \frac{1}{2}r]\right) \times \left(z_2 \in [W_{i2} - \frac{1}{2}r, W_{i2} + \frac{1}{2}r]\right)$ with sampling weight $\hat{f}_H(X_i^S)$, $i = 1, 2, \ldots, n$.

4. Estimate the bandwidth matrix H by the multivariate plug-in estimator of Wand and Jones (1994) and recompute $\hat{f}_H(x)$. Here we should mention that other bandwidth selectors are applicable.

5. Repeat steps 2-4 B (burn-in iterations) $+N$ (additional iterations) times.

6. Discard the B burn-in density estimates and get the final density estimate of $f(x)$ by averaging the remaining N density estimates $\hat{f}_H(x)$ on the evaluation grid G.

The prospective reader may ask how the algorithm fits into existing estimation frameworks. Generally, a popular fitting algorithm for models that depend on latent, unobserved data (the X_i values in our case) is the Expectation Maximization (EM) algorithm. The proposed algorithm is a variant of the classical EM algorithm, namely the Stochastic Expectation Maximization (SEM) algorithm (Celeux et al., 1996). The SEM algorithm works by drawing samples from the conditional distribution $\pi(X_i|W_i)$ creating a pseudo sample of X in each iteration as a replacement of the E-step in the classical EM algorithm where the conditional expectation of X_i given W_i is computed analytically. The classical EM approach would clearly not work for kernel density estimation with rounded data because all the observations within the rectangle around W_i would still be concentrated at a single point, namely the expectation of the conditional distribution of X_i given W_i, computed in the E-Step, leading to spiky estimates of the density. In its original form both the EM and SEM algorithm are used for maximum likelihood estimation in the presence of unobserved variables. However, kernel density estimation is a non-parametric method. We therefore utilize a generalization of the SEM algorithm for the use of surrogates of the likelihood in the M-step (McLachlan and Krishnan, 2007) such that the objective of maximization, i.e. the likelihood, is replaced by the minimization of the AMISE of the kernel density estimator in our case.

The estimator we propose in this paper – hereafter referred to as *GRSST* estimator – allows for estimating the bandwidth matrix H simultaneously with the density. In contrast, for the algorithm proposed by Blower and Kelsall (2002) it is not immediately clear how to estimate H. Blower and Kelsall (2002) suggest using an initial estimate based on the rounded data. Another advantage is that with the proposed algorithm we can get an estimate of the variance induced by the rounding process. This is obtained as a byproduct of the Monte-Carlo process. In particular, standard errors for the density estimates at some arbitrary point can be computed by using the $\hat{f}_H(x)$ produced in each iteration of the algorithm. The algorithm we propose in this paper is also linked to the one proposed by Wang and Wertelecki (2013) in the univariate case. Apart from being derived only for the univariate case, the approach by Wang and Wertelecki (2013) corresponds (in the univariate case) to the method we propose in this paper with $B = 0$ burn-in iterations and $N = 1$ or more sampling steps. However, without a burn-in period no convergence is achieved and final estimates can heavily depend on the pilot estimate. The influence of the burn-in iterations and the sampling steps on the quality of density estimation is evaluated in a simulation study the results of which are included as part of the supporting information. The algorithm is implemented by using function *dbivr* in the **Kernelheaping** R package (Gross, 2015), which is available on CRAN. Additionally, the proposed approach allows for the use of an adaptive bandwidth selection method proposed by Davies et al. (2011) and is implemented in the **sparr** package.

4 Analysis of the Berlin Register of Residents

The benefits of using the proposed multivariate kernel density estimator that accounts for measurement error are illustrated in two applications both of which use the Berlin register data we described in Section 2. The first application aims at estimating the density of the population with migration background in Berlin. The density estimates are compared to the current geographical distribution of advisory centres for migrants in Berlin. The second application aims at estimating the density of the population aged 60 and above in the Berlin area. The density estimates are compared to the current geographical distribution of care homes in the Berlin area.

The analysis is carried out by using the two variables (a) *Migration* and (b) *Age over 60*. The setup of the analysis is as follows: To start with, we impose grids on the geographical space of the Berlin data set with respective grid sizes of 250, 500, 1250, 2000 and 2500 meters. The grid sizes correspond to different degrees of measurement error used for anonymisation purposes. Note that the use of the 2000m by 2000m grids is because these are of similar size to the currently used urban planning areas in Berlin a level at which data is publicly available. Subsequently, we estimate the density of the target population by using the $Naive$ and the proposed $GRSST$ density estimators for each of the grid sizes. We use $B = 5$ and $N = 20$ iterations for the proposed $GRSST$ method in the algorithm presented in Section 3.4. The sensitivity of the density estimators to the size of the dataset, (n) and the effect of the burn-in size, (B) and sample steps (N) is assessed in the supporting information.

The performance of a generic density estimator $\hat{f}(x)$, for example the $Naive$ or the $GRSST$, is typically evaluated by the root mean integrated squared error (RMISE), which is approximated by a Riemann sum over an equally-spaced fine grid,

$$\text{RMISE}(\hat{f}(x)) = \sqrt{E\left(\int (f(x) - \hat{f}(x))^2 dx\right)} \approx \sqrt{\frac{1}{m}\sum_{j=1}^{m}(f(g_j) - \hat{f}(g_j))^2 \delta_g^2}, \tag{5}$$

where m is the number of grid points g_j and δ_g is the gridwidth. For computing the $Naive$ estimator and the $GRSST$ estimator (using the algorithm in Section 3.4) we use a bivariate Gaussian kernel and the plug-in bandwidth selector of Wand and Jones (1994). This is implemented by using the R functions *kde* (kernel density estimation) and *Hpi* (bandwidth selector) provided by the **ks** package (Duong, 2014). The unobserved *true* density $f(x)$ is substituted by the kernel density estimator (1) that uses the original data without rounding with bivariate Gaussian kernel and the plug-in bandwidth selector. This is treated as a benchmark because it is not affected by rounding error. At this point we must mention that the original data is available only to the data host. Hence, for implementing the code with the original data we had to collaborate with staff at the Berlin-Brandenburg Statistics Office. Table 2 shows the goodness of fit in terms of RMISE for the $Naive$ and the proposed density estimators and for different grid sizes. Figures 3

Table 2: Berlin register data: RMISE for *Naive* and *GRSST* multivariate kernel density estimators for different grid sizes (results in units of 10^{-8})

	$r = 250m$		$r = 500m$		$r = 1250m$		$r = 2000m$		$r = 2500m$	
Variable	*Naive*	*GRSST*	*Naive*	*GRSST*	*Naive*	*GRSST*	*Naive*	*GRSST*	*Naive*	*GRSST*
Age above 60	0.66	0.67	1.32	1.27	4.52	2.46	14.08	4.06	23.34	4.66
Migration	0.98	0.97	1.98	1.84	7.33	3.43	22.07	6.12	36.94	6.31

and 4 present kernel density estimation plots for selected grid sizes for *Age over 60* and *Migration* respectively. To start, we note that the proposed estimator outperforms the *Naive* estimator especially for large grid sizes ($\geq$ 1250m). For grid sizes larger or equal to 1250m the *Naive* estimator produces small spikes at the location of the grid points since in this case the probability mass is mostly attributed to the center points of the grid. In contrast, the proposed estimator preserves the fundamental structure of the underlying density. For the largest grid size (2500m), which implies strongly anonymised data, the general shape produced with the proposed estimator is clearly visible. This is not the case with the *Naive* estimator.

Having assessed the performance of both estimators, we now discuss the results of the density estimates in the context of two applications.

Advisory services for population with migration background: Around 950,000 people with migration background from around 190 countries live in the 12 districts in Berlin. The four largest communities consist of approximately 200,000 people with Turkish migration background, around 100,000 people from Russia or from the former Soviet Union and its successor states, approximately 60,000 people of migration background from the successor states in the former Yugoslavia and around 45,000 people of Polish migration background. The history of many migrants started in former West Berlin in the mid-sixties with the recruitment of guest workers. Workers were recruited mainly from Mediterranean countries like Greece, Italy, Yugoslavia or Turkey. In the former East Berlin workers were employed by inter-state agreements from countries like Angola, Poland or Vietnam. From the very beginning Berlin offered advisory services for migrants. For instance, Berlin has a commissioner for integration and migration. This office was established in 1981 and was the first of its kind in Germany. Nowadays, there are specialized advisory service centres that assist people with migration background. The youth migration services provide advice to young adults and teenagers of migration background. In addition, Berlin has in total 32 advisory service centres for adults. In these centres migrants can receive support and personal consultation directly that will assist with their integration. For example, people receive support with finding appropriate child care facilities. To secure an appropriate level of support it is important to establish advisory centres where mostly needed. The left panel of Figure 5 shows the estimated densities of the population with migration background in Berlin. The blue points represent the 32 advisory service centres for adults. The plot on the top panel shows the density estimates produced by using the original data and the exact address coordinates, which are not publicly available.

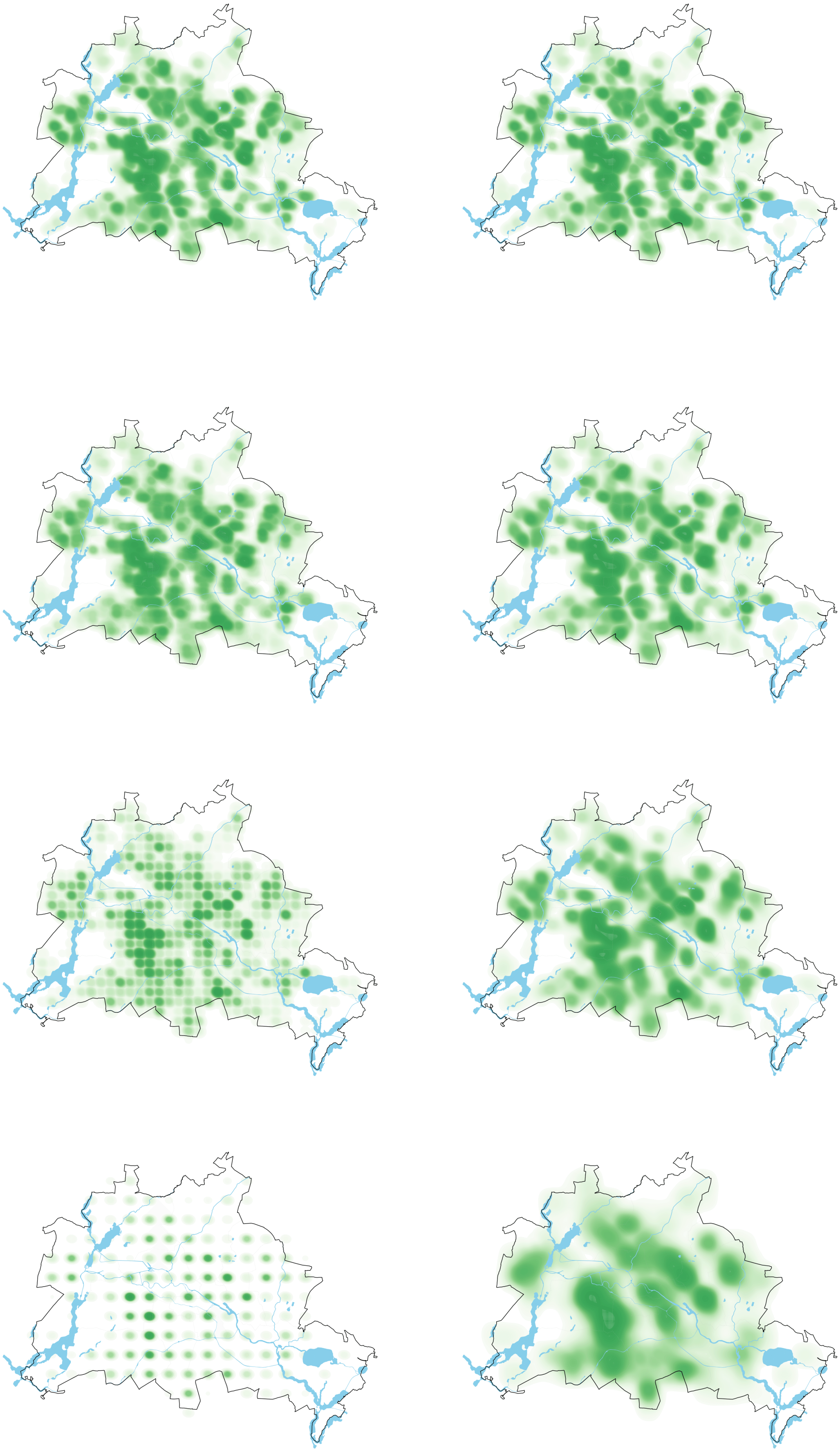

Figure 3: Density estimates of population aged 60 and above: *Naive* (left panel) and *GRSST* estimators (right panel) with rounding step sizes of 0 (original data), 500, 1250 and 2500 m (top down).

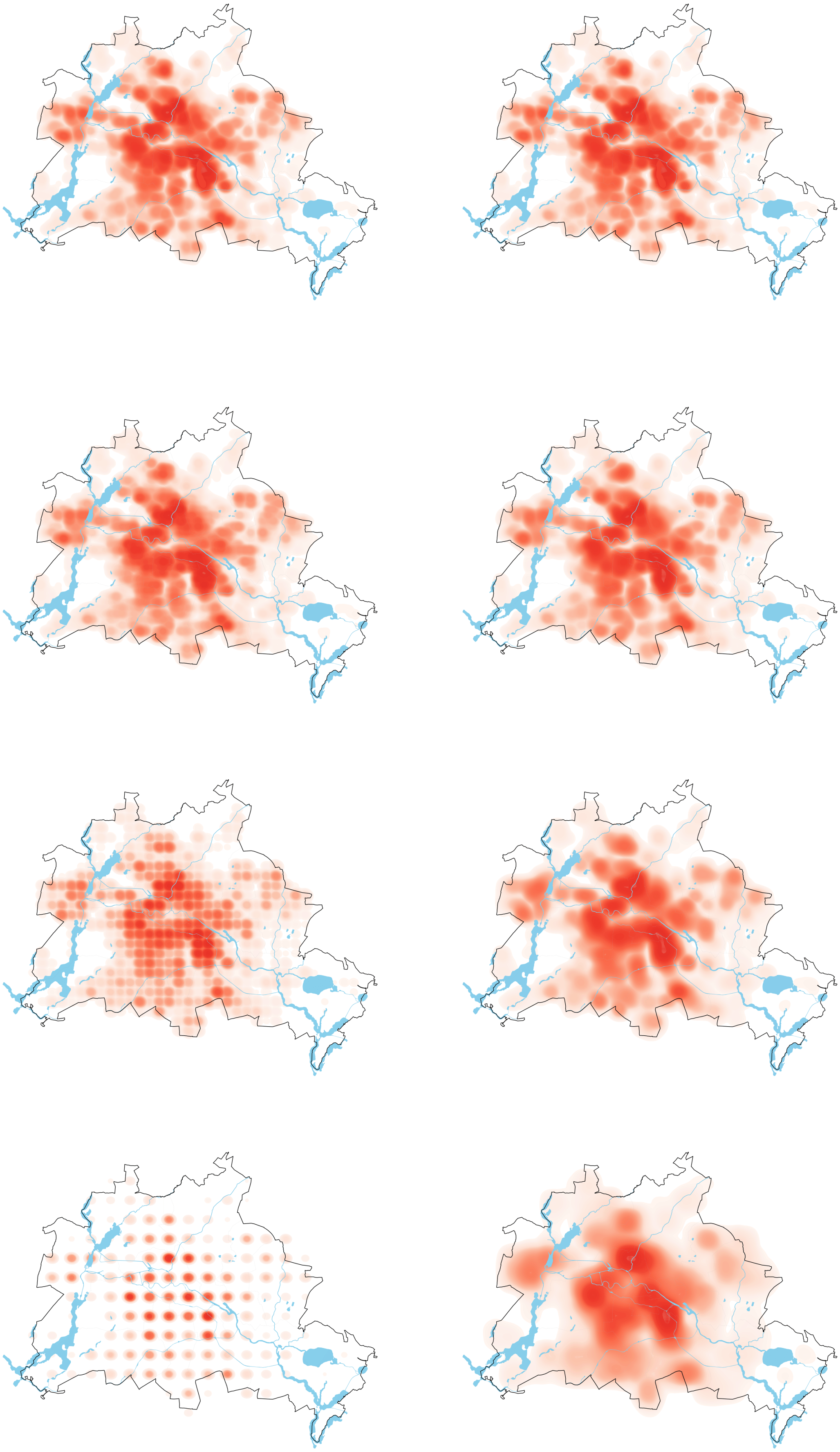

Figure 4: Density estimates of population with migration background: *Naive* (left panel) and *GRSST* estimators (right panel) with rounding step sizes of 0 (original data), 500, 1250 and 2500 m (top down).

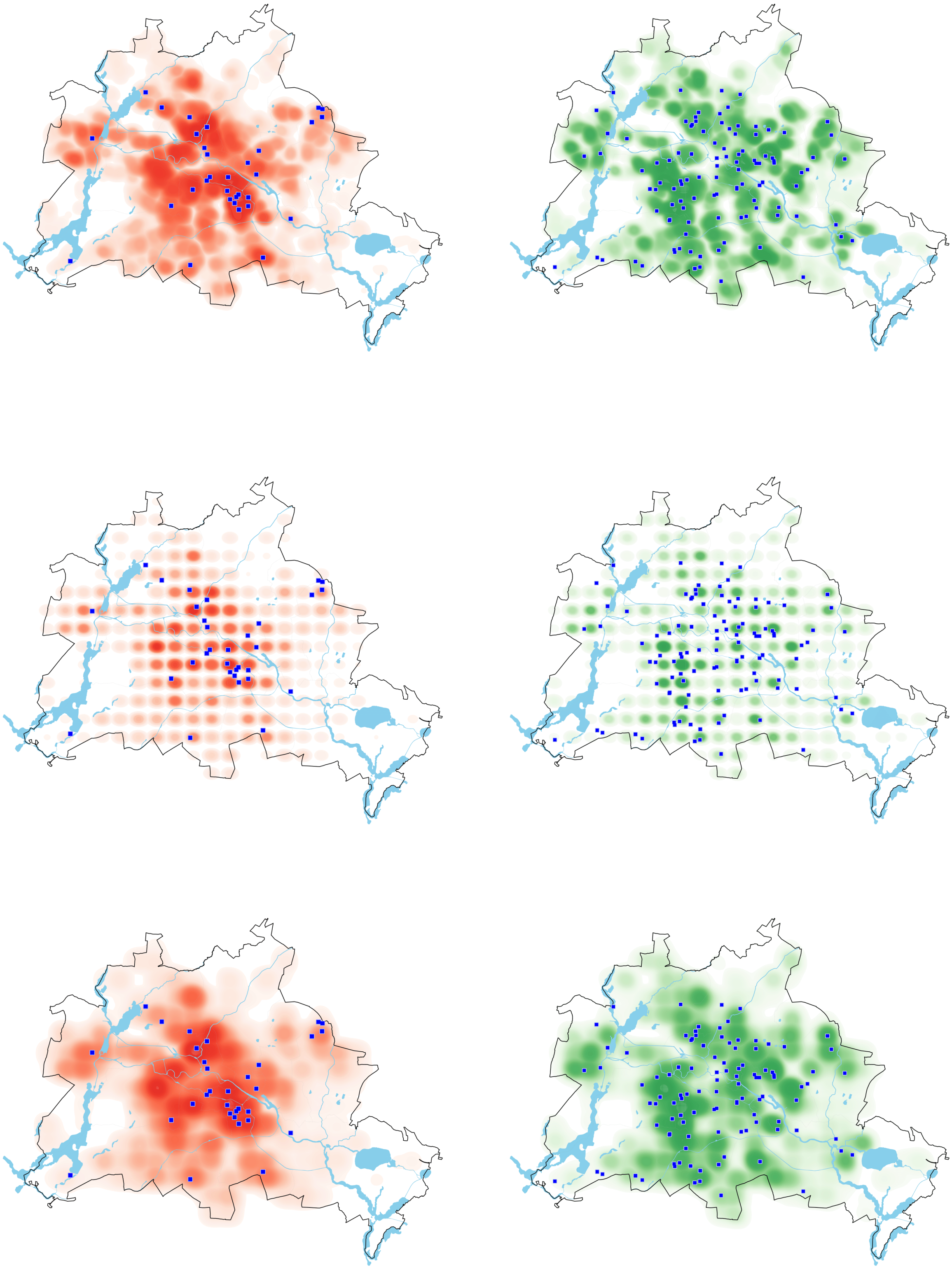

Figure 5: Migration background (left panel) and Age above 60 (right panel) for the original data, *Naive* method and *GRSST* method (top down) for rounding step size of 2000 m including points of interest. Blue points indicate migrant advisory centers and retirement houses respectively.

The plots in the middle and at the bottom present density estimates produced by using the *Naive* and the *GRSST* density estimators with a rounding step size of 2000m. The choice of 2000m times 2000m grids is because these are of similar size to the currently used urban planning areas in Berlin. The estimates based on the original data in Figure 5 show that the density of populations with migration background varies by Berlin districts. The estimated density is particularly high in the former West-Berlin districts of Wedding (in the north), Neukölln (in the south-east), Kreuzberg (in the center to south) and Schöneberg (in the south). Friedrichshain and Prenzlauer Berg (in the north-east), show a lower estimated density of population with migration backgrounds.

The spatial distribution of advisory centres cover populations in the centre and north of Berlin quite well. However, there are some hotspots for example, in the western and south-west parts (Charlottenburg or Moabit) or in the very northern parts (Märkisches Viertel) of Berlin, with a high density estimate of ethnic minority populations but without any advisory service centres. The commentary on the first map above depends on precise geo-coded addresses which are not publicly available. The second and third maps show the density estimates based on the rounded data. The density plot obtained by using the *Naive* estimator (plot in the middle in Figure 5) produces spikes at the center of the grids. In contrast, the proposed estimator produces a map (plot at the bottom in Figure 5) that is able to preserve the fundamental density structure of the original data. Hence, the commentary we produced by looking at the map of the original data holds also true for the map of density estimates produced by using the proposed multivariate kernel density estimator that accounts for measurement error. In addition, the proposed density estimator produces more precise density estimates than the *Naive* one (see Table 2). Local authorities should prefer the density estimates produced by the proposed estimator, to the one produced by the *Naive* estimator, for making informed decisions.

Care for the elderly: Life expectancy in Germany has improved due to advances in medical research. This leads to a change in the demographic structure with an increasing number of old-aged people. Approximately 860,000 individuals aged 60 and above live in Berlin. It is projected that by 2030 the average age of Berlin's population will increase from 42.5 years (in 2007) to 45.3 years and roughly every third citizen of Berlin will be 60 years or older. With increasing age the prevalence of diseases and functional restraints, which are often chronic and irreversible, rises as well (Saß et al., 2009). In 2012, 58.3% of German women and 55.3% of German men suffered from at least one chronic disease (Robert Koch Institute, 2014). According to the World Health Organization (2005), the prevalence and incidence of various chronic diseases, such as cardiovascular diseases, cancer, diabetes mellitus, dementia or respiratory problems, is predicted to increase in the next years. For this reason older people are more likely to need help in their daily life and will increasingly depend on care. According to the nursing care insurance in 2011 there were roughly 117,500 care-dependent people in Berlin. In order to support the increasing elderly population it is necessary to offer high-quality medical and social community structures of care that are close to the people's place of residence. This is important because elderly people tend to feel connected to their neighbourhood. These structures consist of:

- Neighborhood centers: These are combinations of accessible living, residential care homes and social/cultural centres with neighbourhood cafes, which are suitable for senior citizens. Such structures offer elderly people with or without care dependency the opportunity to live actively within the community until old age.
- Foster ambulatory care: These are home care nursing services that enable care-dependent people to live at home.
- Networked care: The different forms of care systems (e.g., ambulatory care, semi-residential care or impatient care) need to be more strongly interconnected than they currently are. This will offer more choices for older people for example, live at home with ambulatory care but have the opportunity to change to semi-residential or impatient care near to the place they live.

In order to improve such services for the city of Berlin it is necessary to have an accurate picture about the distribution of the elderly population in Berlin. The right panel of Figure 5 shows the density estimates for the population aged 60 years or above. The blue points represent 108 retirement homes in Berlin. The location of these points was extracted by using Google Maps. The plot on the top panel indicates the density estimates based on the original data with the exact address coordinates, which are not publicly available. The plots in the middle and at the bottom present the density estimates by using the $Naive$ and the proposed density estimators with a rounding step size of 2000m. The supply of retirement houses is particularly good in the center of Berlin. However, locations for future expansion of retirement houses and other support structures can be identified. For instance, there are some hotspot areas in the north (Reinickendorf and especially Märkisches Viertel) or in the south-east (Gropiusstadt) with a high density estimate of the population over 60 but without retirement homes. As in the first application,the proposed estimator (plot at the bottom in Figure 5) preserves the structure of the density of the population over 60 years despite the presence of measurement error in the available data and offers more precise estimates. Hence, the use of the proposed estimator may enable local authorities and other organisations to make sound strategic decisions regarding the best places for investigating in creating infrastructure for social care without requiring access to exact geo-referenced data. A more refined analysis of the Berlin register data could consider the use of local bandwidths as opposed to a global bandwidth. This is possible by using the R package that has been written for implementing the methodology we propose in this paper. Nevertheless, use of local bandwidths can increase significantly the computational time.

5 Simulation Study

In this section we present results from a Monte-Carlo simulation study that was conducted for evaluating the performance of the proposed multivariate kernel density estimator we presented in Section 3. The objective of this simulation study is to investigate the ability of the proposed methodology to account for measurement error, under different scenarios for the intensity of the measurement error process, and hence provide more precise estimates than *Naive* kernel density estimation that disregards measurement error. The proposed estimator is further compared to the estimator proposed by Delaigle (2007). Finally, the sensitivity of the proposed method in relation to the size of the data (n), to the burn-in size (B) and sample steps (N) used in the $GRSST$ algorithm is evaluated and the results are provided as part of the supporting information.

The simulation data is generated under different bivariate normal distributions. Three scenarios, denoted by A, B and C, are considered. Under Scenario A data is generated by using a bivariate standard normal distribution,

$$f_A(x) = \phi(x|\mu, \Sigma),$$

where $\phi(x|\mu, \Sigma)$ denotes a multivariate normal density with mean μ and variance-covariance matrix Σ given by,

$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad , \quad \Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Under Scenario B data is generated by using a mixture of three uncorrelated bivariate normal distributions,

$$f_B(x) = \frac{1}{3}\phi(x|\mu_1, \Sigma_1) + \frac{1}{3}\phi(x|\mu_2, \Sigma_2) + \frac{1}{3}\phi(x|\mu_3, \Sigma_3),$$

with

$$\mu_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mu_2 = \begin{pmatrix} 5 \\ 3 \end{pmatrix}, \mu_3 = \begin{pmatrix} -4 \\ 1 \end{pmatrix}, \Sigma_1 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \Sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}.$$

Finally, under Scenario C data is generated by using a mixture of three correlated normal distributions with

$$\mu_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mu_2 = \begin{pmatrix} 5 \\ 3 \end{pmatrix}, \mu_3 = \begin{pmatrix} -4 \\ 1 \end{pmatrix}, \Sigma_1 = \begin{pmatrix} 4 & 3 \\ 3 & 4 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 3 & 0.5 \\ 0.5 & 1 \end{pmatrix}, \Sigma_3 = \begin{pmatrix} 5 & 4 \\ 4 & 6 \end{pmatrix}.$$

The corresponding density contours under the three scenarios are shown in Figure 6. The use of bivariate distributions is motivated by the fact that our application data in Section 4 is bivariate. The use of Gaussian distributions for generating the simulation data follows Zhang et al. (2006) and Zougab et al. (2014).

For each scenario we generate a dataset S_0 of size $n = 500$ from the corresponding

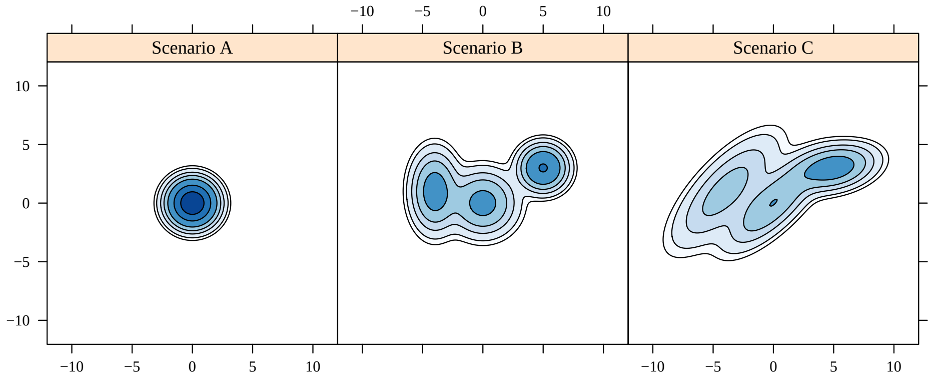

Figure 6: Contour plots of the simulated data under the three simulation scenarios.

distribution $f_A(x)$, $f_B(x)$ or $f_C(x)$. The dataset S_0 includes the exact x- and y-coordinates. For introducing measurement error via rounding of the coordinates, we define a grid for the x- and y-coordinates ranging from -10 to 10 with gridwidth according to rounding values $r = 0.75, 1.5$ and 2.25. For a formal definition of r and the rounding process we refer to Section 3.2. We denote the dataset after rounding by S_r. Figure 7 shows the different scenarios for the rounding process for a specific dataset under Scenario B. The size of the points represents the number of points at a specific rounding tick.

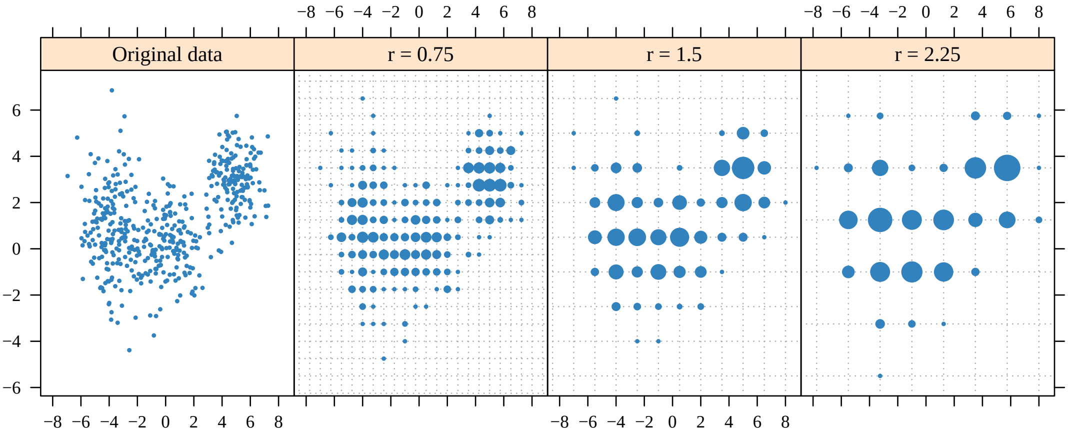

Figure 7: Scenario B: Rounding procedure for a specific dataset.

By using S_r, we estimate the density with three methods: a) *Naive*: a standard kernel density estimator that ignores measurement error, b) *GRSST*: This is the proposed SEM estimator with $B = 5$ burn-in and $N = 20$ sample steps and c) Delaigle: this is the estimator presented in Delaigle (2007). As in Section 4, for computing the *Naive* and *GRSST* estimators we use a bivariate Gaussian kernel and a plug-in bandwidth selector. The density of the original data S_0 ($r = 0$ in Table 3) is estimated by using function *kde* (kernel density estimation) with a bivariate Gaussian kernel and a plug-in bandwidth

Table 3: Mean RMISE for different grid sizes (r) and scenarios. Corresponding standard errors of the RMISE in parentheses.

	$r = 0$	$r = 0.75$			$r = 1.5$			$r = 2.25$		
	Original	*Naive*	*GRSST*	Delaigle	*Naive*	*GRSST*	Delaigle	*Naive*	*GRSST*	Delaigle
Scenario A	0.205 (0.026)	0.238 (0.029)	0.239 (0.031)	0.301 (0.027)	3.952 (0.301)	0.242 (0.030)	0.887 (0.034)	4.917 (0.248)	0.568 (0.045)	1.113 (0.056)
Scenario B	0.162 (0.016)	0.172 (0.017)	0.170 (0.016)	0.328 (0.019)	0.380 (0.033)	0.183 (0.018)	0.272 (0.013)	0.679 (0.043)	0.256 (0.016)	0.390 (0.014)
Scenario C	0.119 (0.012)	0.125 (0.013)	0.121 (0.012)	0.268 (0.015)	0.147 (0.013)	0.131 (0.013)	0.181 (0.009)	0.351 (0.034)	0.152 (0.014)	0.172 (0.012)

selector. The density estimates of the original data are treated as a *benchmark* because S_0 is not affected by rounding error. The simulation steps (generation of a dataset, rounding of the coordinates and the density estimation) are independently repeated 500 times for each scenario.

In Table 3 we compare the performance of the $Naive$, the $GRSST$ and the Delaigle density estimators in the three scenarios. The first column of Table 3 shows the means and the standard deviations of the RMISE over 500 Monte-Carlo replications of the *benchmark* case i.e. in the absence of rounding error ($r = 0$). Note that in the definition of the RMISE in (5) $f(x)$ denotes now the underlying true density, $f_A(x), f_B(x)$ or $f_C(x)$ respectively.

For the scenarios with small rounding errors ($r = 0.75$) we observe that the $Naive$ and the $GRSST$ density estimators perform similarly and both methods have RMISE which is comparable to the RMISE under the *benchmark* scenario. The Delaigle estimator reveals a higher RMISE compared to the two other approaches. Data providers may be keen, however, to introduce more severe measurement error to the data for ensuring confidentiality. For such scenarios ($r = 1.5$ and $r = 2.25$) the $GRSST$ density estimator clearly outperforms the $Naive$ estimator. The Delaigle estimator performs better than the $Naive$ estimator but worse compared to the $GRSST$ estimator. It is notable that the $Naive$ estimator performs very poorly especially for $r = 1.5$ and $r = 2.25$ in the case of a bivariate standard normal distribution (Scenario A). Presumably this is due to the small variance of the underlying density we are trying to estimate in Scenario A such that discretizing for given rounding values has a much more pronounced effect. For this reason we also tested a bivariate normal distribution with a larger variance. The results for the $Naive$ method become more stable but the $GRSST$ estimator still performs better. Figure 8 shows contour plots of a particular simulation run under Scenario B for the $Naive$ and $GRSST$ estimators. It appears that, unlike the $Naive$, the $GRSST$ density estimator is able the maintain the underlying structure of the density for different rounding levels. Contour plots under Scenarios A and C (provided as part of the supporting information) confirm this finding. The anisotropic pattern for the $Naive$ estimator ($r = 1.5$ and $r = 2.25$) is caused by a larger bandwidth in x-direction than in y-direction. This bandwidth is chosen by the plug-in bandwidth matrix selector of (Wand and Jones (1994).

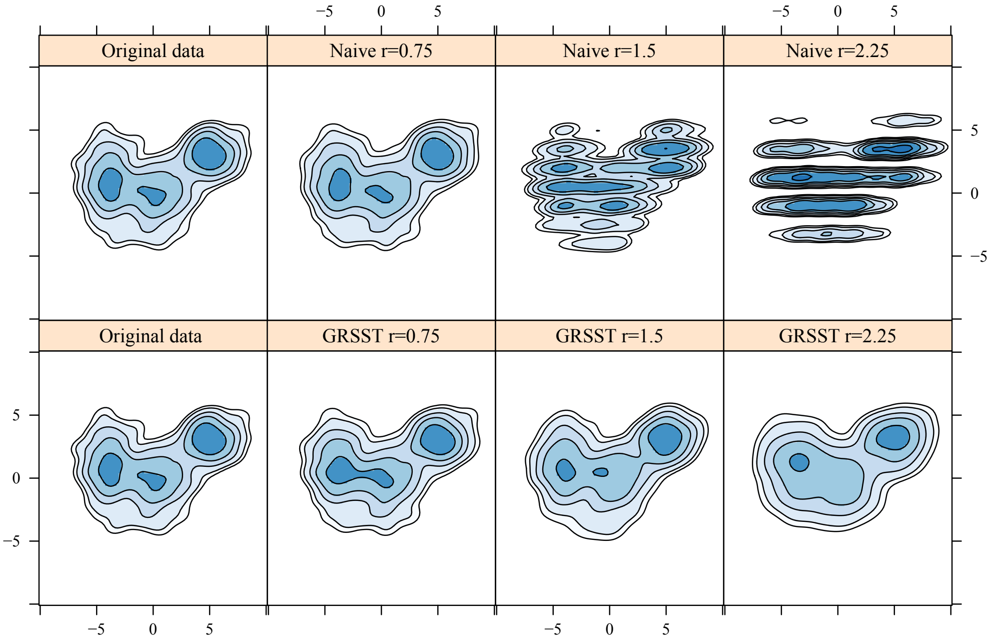

Figure 8: Scenario B: Contour plots of *Naive* estimator (upper panel) and *GRSST* estimator (lower panel), for grid size $r = 0.75, 1.5, 2.25$ (left to right). The original data scenario ($r = 0$) is used as the benchmark.

6 Discussion

Precise geo-coded data is rarely available due to confidentiality constraints. The paper proposes methodology for deriving density estimates of populations of interest in the presence of rounding in the geographical coordinates used for disclosure control. The proposed methodology is motivated by reversing the measurement error process by combining a measurement error model with kernel density estimation. The method is straightforward to implement and works for different dimensions, symmetric as well as asymmetric kernel types and bandwidth selection methods. The use of the proposed methodology is facilitated by the availability of function *dbivr* in the R package **Kernelheaping** available on CRAN (Gross, 2015). As we demonstrated with the analysis of the Berlin register data the proposed method can offer considerably deeper insights, compared to a *Naive* estimator that disregards the measurement error process, to data analysts about the density of target populations within an area of interest. The structure preserving property of the proposed method is particularly attractive when working with data that has been subjected to disclosure control via the introduction of measurement error. In addition, the paper provides some first indications on how to set the grid-lengths for geo-coding in the Berlin register of residents such that a data analyst is able to derive precise density estimates. At the same time working with the data host for deciding the grid-lengths is crucial for ensuring confidentiality.

Further work could extend the proposed approach to different geographical masking or anonymisation methods including for example the use of Gaussian errors added to the original geographic coordinates. With minor adaptions to the algorithm direct use of arbitrary demarcation shapes like the LORs instead of the grid-structure induced by rounding is possible for obtaining smooth density estimates as well. The proposed method can be further generalized for application to data with varying degree of rounding (*heaping*) occurring, for example, in self-reported survey data (Pudney, 2008). Finally, one idea for further work is to explore the application of the proposed methodology for generating synthetic geo-coded data based on anonymised data sets with rounding errors.

Acknowledgments

The authors gratefully acknowledge the data access and support of the Research Data Centre of the Statistical Office for Berlin-Brandenburg and are indebted to the Editor, Associate Editor and two referees for comments that significantly improved the paper.

References

Acevedo-Garcia, D., K. A. Lochner, T. L. Osypuk, and S. V. Subramanian (2003). Future directions in residential segregation and health research: a multilevel approach. *American Journal of Public Health 93*(2), 215–21.

Armstrong, M. P., G. Rushton, and D. L. Zimmerman (1999). Geographically masking health data to preserve confidentiality. *Statistics in Medicine 18*(5), 497–525.

Berkson, J. (1950). Are there two regressions? *Journal of the American Statistical Association 45*(250), 164–180.

Blower, G. and J. E. Kelsall (2002). Nonlinear kernel density estimation for binned data: convergence in entropy. *Bernoulli 8*(4), 423–449.

Card, D. and J. Rothstein (2007, December). Racial segregation and the black-white test score gap. *Journal of Public Economics 91*(11-12), 2158–2184.

Carroll, R., D. Ruppert, L. Stefanski, and C. Crainiceanu (2010). *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition.* Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

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

Davies, T. M., M. L. Hazelton, and J. C. Marshall (2011). sparr: Analyzing spatial relative risk using fixed and adaptive kernel density estimation in R. *Journal of Statistical Software 39*(1), 1–14.

Delaigle, A. (2007). Nonparametric density estimation from data with a mixture of berkson and classical errors. *Canadian Journal of Statistics 35*(1), 89–104.

Delaigle, A. (2014). Nonparametric kernel methods with errors-in-variables: Constructing estimators, computing them, and avoiding common mistakes. *Australian & New Zealand Journal of Statistics 56*(2), 105–124.

Delaigle, A. and I. Gijbels (2004). Practical bandwidth selection in deconvolution kernel density estimation. *Computational Statistics & Data Analysis 45*(2), 249–267.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, 1–38.

Destatis (2009). Germany's population by 2060 - results of the 12th coordinated population projection.

Duong, T. (2014). *ks: Kernel smoothing.* R package version 1.9.0.

Escobar, M. D. and M. West (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association 90*(430), 577–588.

Fan, J. et al. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics 19*(3), 1257–1272.

Feuerverger, A., P. T. Kim, and J. Sun (2008). On optimal uniform deconvolution. *Journal of Statistical Theory and Practice 2*(3), 433–451.

Fuller, W. (2009). *Measurement Error Models.* Wiley Series in Probability and Statistics. Wiley.

Gelfand, A. E., A. Kottas, and S. N. MacEachern (2005). Bayesian nonparametric spatial modeling with dirichlet process mixing. *Journal of the American Statistical Association 100*(471), 1021–1035.

Gorr, W., M. Johnson, and S. Roehrig (2001). Spatial decision support system for home-delivered services. *Journal of Geographical Systems 3*, 181–197.

Gross, M. (2015). *Kernelheaping: Kernel Density Estimation for Heaped and Rounded Data.* R package version 1.0.

Härdle, W. and D. W. Scott (1992). Smoothing by weighted averaging of rounded points. *Computational Statistics 7*, 97–128.

Izenman, A. J. (1991). Recent developments in nonparametric density estimation. *Journal of the American Statistical Association 86*(413), pp. 205–224.

Jones, M. C., J. S. Marron, and S. J. Sheather (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association 91*(433), 401–407.

Kwan, M.-P., I. Casas, and B. C. Schmitz (2004). Protection of geoprivacy and accuracy of spatial information: How effective are geographical masks? *Cartographica: The International Journal for Geographic Information and Geovisualization 39*(2), 15–28.

Long, J. P., N. E. Karoui, and J. A. Rice (2014). Kernel density estimation with berkson error. *arXiv preprint arXiv:1401.3362*.

Marron, J. S. (1987). A comparison of cross-validation techniques in density estimation. *The Annals of Statistics 15*(1), 152–162.

McLachlan, G. and T. Krishnan (2007). *The EM algorithm and extensions*, Volume 382. John Wiley & Sons, New York.

Minnotte, M. C. (1998). Achieving higher-order convergence rates for density estimation with binned data. *Journal of the American Statistical Association 93*(442), 663–672.

Ozonoff, A., C. Jeffery, J. Manjourides, L. F. White, and M. Pagano (2007). Effect of spatial resolution on cluster detection: a simulation study. *International Journal of Health Geographics 6*(1), 1–7.

Peterson, R. D., L. J. Krivo, and C. R. Browning (2008). Segregation and race/ethnic inequality in crime: New directions. In F. T. Cullen, J. Vright, and K. Blevins (Eds.), *Taking stock : the status of criminological theory.* New Brunswick, NJ: Transaction.

Pudney, S. (2008). Heaping and leaping: Survey response behaviour and the dynamics of self-reported consumption expenditure. Technical report, ISER Working Paper Series.

Robert Koch Institute (2014). Beiträge zur Gesundheitsberichterstattung des Bundes - Daten und Fakten: Ergebnisse der Studie ”Gesundheit in Deutschland aktuell 2009”. `http://www.rki.de/DE/Content/Gesundheitsmonitoring/Gesundheitsberichterstattung/GBEDownloadsB/GEDA09.pdf?__blob=publicationFile`.

Rushton, G., M. Armstrong, J. Gittler, B. Greene, C. Pavlik, M. West, and D. Zimmerman (2007). *Geocoding Health Data: The Use of Geographic Codes in Cancer Prevention and Control, Research and Practice.* Taylor & Francis.

Saß, A., S. Wurm, and T. Ziese (2009). Somatische und psychische Gesundheit. In *Beiträge zur Gesundheitsberichterstattung des Bundes - Gesundheit und Krankheit im Alter*, pp. 31–61. K. Böhm and C. Tesch-Römer and T. Ziese.

Scott, D. W. and S. J. Sheather (1985). Kernel density estimation with binned data. *Communications in Statistics - Theory and Methods 14*(6), 1353–1359.

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

Stefanski, L. A. and R. J. Carroll (1990). Deconvolving kernel density estimators. *Statistics: A Journal of Theoretical and Applied Statistics 21*(2), 169–184.

VanWey, L. K., R. R. Rindfuss, M. P. Gutmann, B. Entwisle, and D. L. Balk (2005). Confidentiality and spatially explicit data: Concerns and challenges. *Proceedings of the National Academy of Sciences of the United States of America 102*(43), 15337–15342.

Verma, I. (2014). Planning for aging neighborhoods. In *Proceedings of the 6th Annual Architectural Research Symposium in Finland*, Number 6 in Annual Architectural Symposium in Finland.

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

Wang, B. and W. Wertelecki (2013). Density estimation for data with rounding errors. *Computational Statistics & Data Analysis 65*, 4–12.

World Health Organization (2005). Preventing chronic diseases: a vital investment. `http://whqlibdoc.who.int/publications/2005/9241563001_eng.pdf`.

Xu, S. (2014). Asymmetric kernel density estimation based on grouped data with applications to loss model. *Communications in Statistics-Simulation and Computation 43*(3), 657–672.

Zhang, C.-H. (1990). Fourier methods for estimating mixing densities and distributions. *The Annals of Statistics 18*(2), 806–831.

Zhang, X., M. L. King, and R. J. Hyndman (2006). A bayesian approach to bandwidth selection for multivariate kernel density estimation. *Computational Statistics & Data Analysis 50*(11), 3009–3031.

Zougab, N., S. Adjabi, and C. Kokonendji (2014). Bayesian estimation of adaptive bandwidth matrices in multivariate kernel density estimation. *Computational Statistics & Data Analysis 75*(11), 28–38.