Freie Universität Berlin

Bayesian Kernel Density Estimation applied to sensitive Geo-Coded Data of Berlin

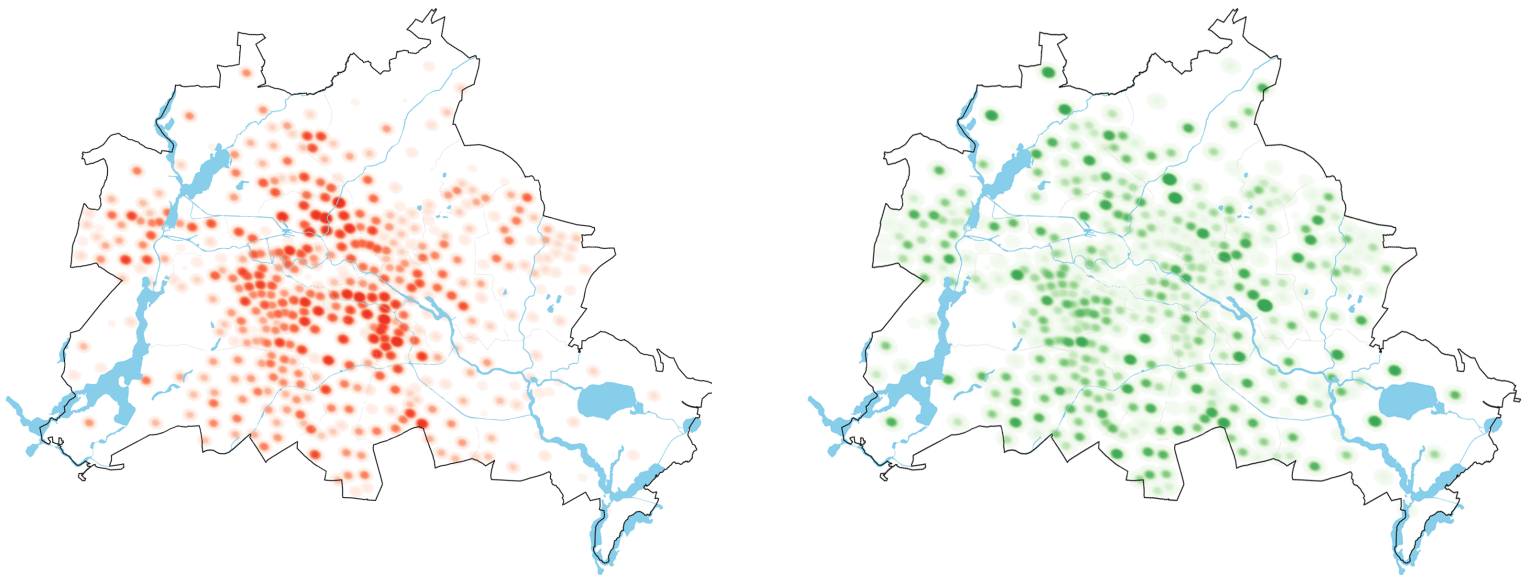
Ulrich Rendtel
Marcus Groß,, Timo Schmid, Sebastian Schmon and Nikos Tzavidis

Banocoss Conference 2015 Helsinki
August 2015

Freie Universität Berlin

Motivation - Density estimates in Berlin

Ethnic minority background in Berlin (left map) and **aged people** in Berlin (right map) based on the publicly available data.



$\Longrightarrow$ Fundamental density **structure is not preserved**

Freie Universität Berlin

Motivation

- Modern systems of **official statistics require** the timely estimation of area-specific densities of sub-populations.
- Estimates should be based on **precise geo-coded information** - **hardly available** due to confidentiality constraints.
- A version of the Berlin register data is publicly available including aggregates for the 447 urban planning areas (LOR) - Fundamental density **structure is not preserved**.
- **Research question**: Can we derive precise density estimates of sub-groups by using data that has been subjected to disclosure control via aggregation or rounding of the geographic coordinates?

Freie Universität Berlin

Table of contents

Freie Universität Berlin

Multivariate kernel density estimation

For a sample of bivariate data $X = (X_1, ..., X_n)$ from a density f the **kernel density** estimate is **defined** as

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

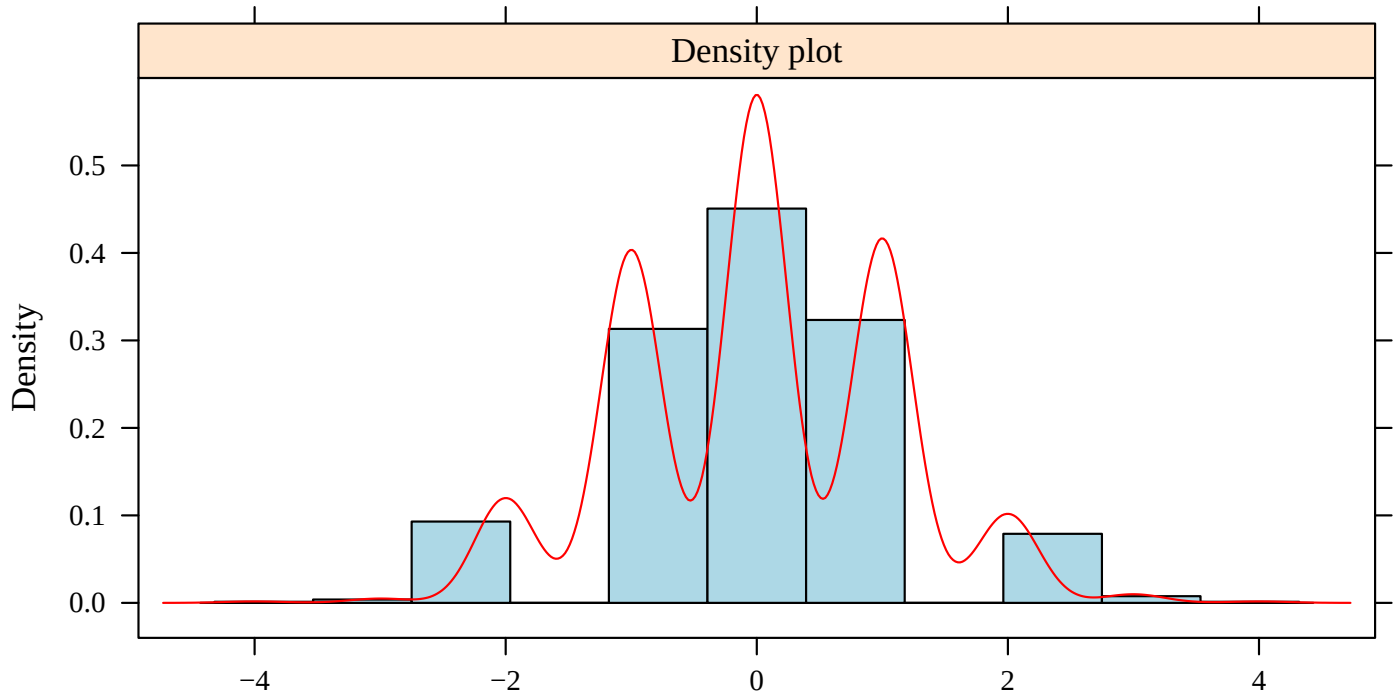
- $K()$ is a two dimensional kernel function and H is a bandwidth matrix.
- Through the anonymisation process only the rounded values W_i are available.

$\Longrightarrow$ Does the anonymisation process effect the density estimates?

Freie Universität Berlin

Effect of the anonymisation process

Univariate case:

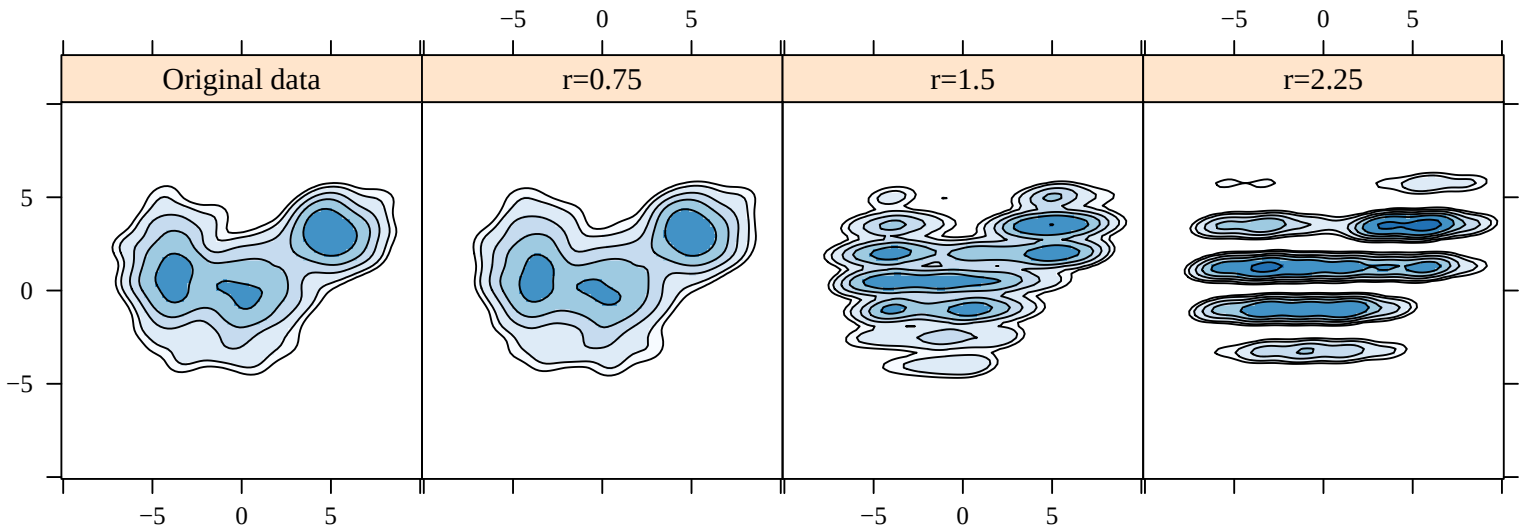


- Kernel density estimator applied on data generated under normality (values rounded to the next integer).

Freie Universität Berlin

Effect of the anonymisation process

Multivariate case:



- Multivariate kernel density estimator applied on data (mixture of three uncorrelated bivariate normal distributions).

Idea:
Regard the anonymisation process as a **Berkson measurement error** on the true values X_i (Berkson, 1950).

Freie Universität Berlin

Rounding and kernel density estimation

Berkson measurement error model (Berkson, 1950):

- 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,

$$[W_{i1} - \frac{1}{2}\mathrm{r}, W_{i1} + \frac{1}{2}\mathrm{r}] \times [W_{i2} - \frac{1}{2}\mathrm{r}, W_{i2} + \frac{1}{2}\mathrm{r}],$$

  r denotes the rounding parameter.

- Can be translated into a Berkson measurement error model with uniformly distributed measurement error $U_i = (U_{i1}, U_{i2})$, $U_{i1}, U_{i2} \sim Unif(-\frac{1}{2}r, \frac{1}{2}r)$ and U_{i1}, U_{i2} independent of W_{i1} and W_{i2} such that,

$$X_{i1} = W_{i1} + U_{i1}, \ \ i = 1, 2, .., n$$

$$X_{i2} = W_{i2} + U_{i2}, \ \ i = 1, 2, .., n.$$

Freie Universität Berlin

Bayesian measurement error model

- Treat the unknown true values X_i as latent variables
- W_i only depends on X_i, the Likelihood can be split in two parts (measurement error model and observation model).

Posterior distribution by using a hierarchical model:

$$\pi(X, H|W) \propto \underbrace{\pi(W|X) \times \pi(X|H)}_{\text{Likelihood}} \times \underbrace{\pi(H)}_{\text{Prior}}$$

where $\pi(W|X)$ (measurement error model) is the indicator function and $\pi(X|H)$ (observation model) is defined by the leave-one-out kernel density estimator (Zhang et al., 2006)

Freie Universität Berlin

Computational details

Iterative MCMC-type algorithm:

1. Get a pilot estimate of f_X by setting H to $\begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}$, where I is a sufficiently *large* value such that no rounding spikes occur.
2. **Evaluate the density** estimate $\hat{f}_X$ on an **equally-spaced fine grid** $G = \tilde{x_1} \times \tilde{x_2}$.
3. X_i **is repeatedly drawn** from the square of side length r around W_i according to the current density estimate $\hat{f}_X$.
4. **Estimate the bandwidth matrix** H by the multivariate plug-in estimator of Wand and Jones (1994) and recompute $\hat{f}_X$.
5. Repeat steps 2-4 B (burn-in iterations) + N (additional iterations) times.
6. Discard the burn-in samples and get **final estimate of** f_X **by averaging** over the remaining samples.

Freie Universität Berlin

Table of contents

Freie Universität Berlin

Empirical Studies - Aims

Model-based simulations:

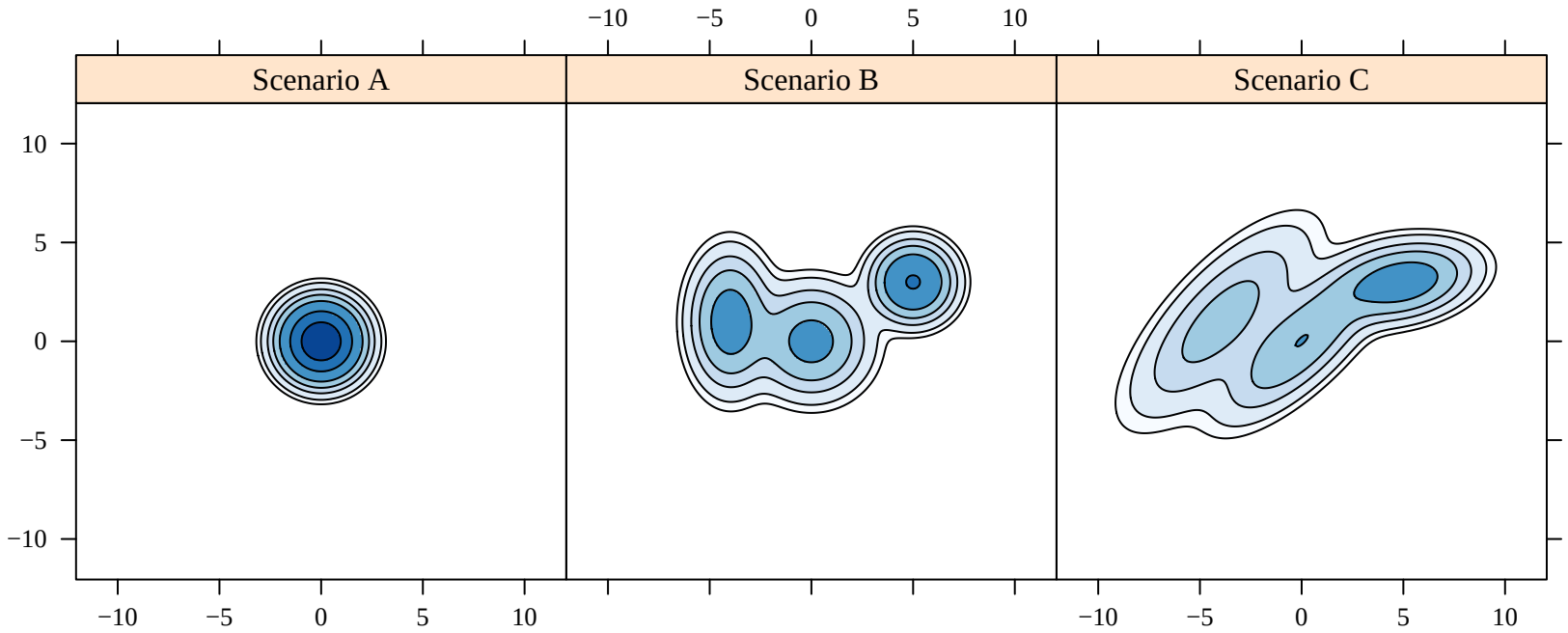
- Investigate the ability of the proposed methodology to account for measurement error, under different scenarios for the intensity of the measurement error process.

Application to Berlin Register of Residents:

- Evaluate the performance the density estimators under a more realistic setting.
- Discuss the results of the density estimates in the context of two applications.

Freie Universität Berlin

Simulation setup

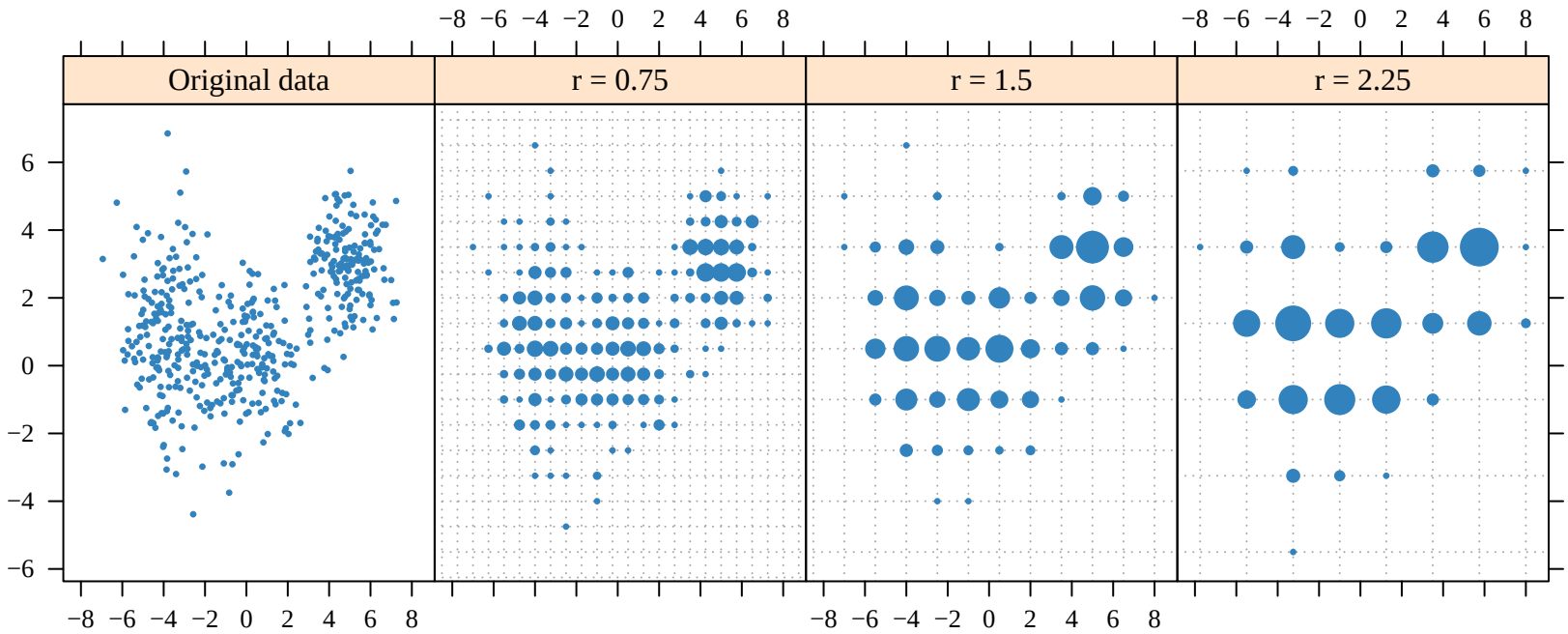


Contour plots of the simulated data under the three simulation scenarios.

- For each scenario we generate a dataset S_0 of size $n = 500$ from the corresponding distribution.
- The dataset S_0 includes the exact x- and y-coordinates.

Freie Universität Berlin

Simulation setup



Scenario B: Rounding procedure for a specific dataset.

- 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.
- The size of the points represents the number of points at a specific rounding tick.

Freie Universität Berlin

Simulation setup

We estimate the density with two methods:

- *Naive*: a standard kernel density estimator that ignores measurement error.
- MCMC: the proposed Bayesian kernel density estimator.

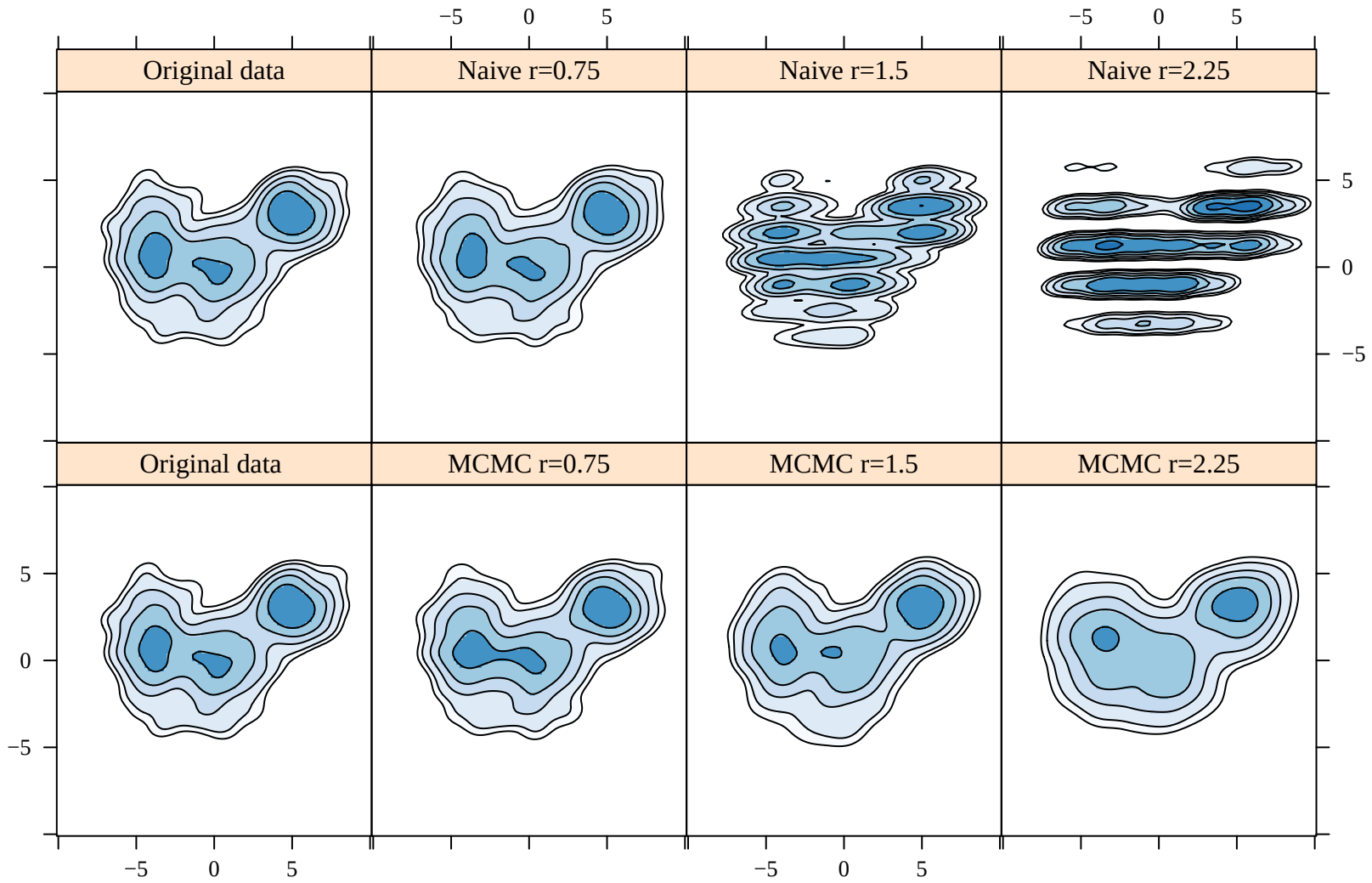
The performance of the density estimates $\hat{f}$ is assessed by the root integrated mean squared error (RMISE):

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

where f denotes the underlying true density, m is the number of grid points g_j and δ_g is the gridwidth.

Freie Universität Berlin

Simulation results: Scenario B



Freie Universität Berlin

Simulation results

Mean RMISE for different grid sizes (r) and scenarios based on 500 Monte-Carlo replications.

	r= 0	r= 0.75		r= 1.5		r= 2.25	
	Original	*Naive*	MCMC	*Naive*	MCMC	*Naive*	MCMC
Scenario A	0.205	0.238	0.239	3.952	0.242	4.917	0.568
Scenario B	0.162	0.172	0.170	0.380	0.183	0.679	0.256
Scenario C	0.119	0.125	0.121	0.147	0.131	0.351	0.152

- For the scenarios with small rounding errors the *Naive* and the MCMC density estimators perform similarly.
- For the scenarios with more severe measurement error (r=1.5 and r=2.25) the MCMC density estimator clearly outperforms the *Naive* estimator.

Freie Universität Berlin

Table of contents

Freie Universität Berlin

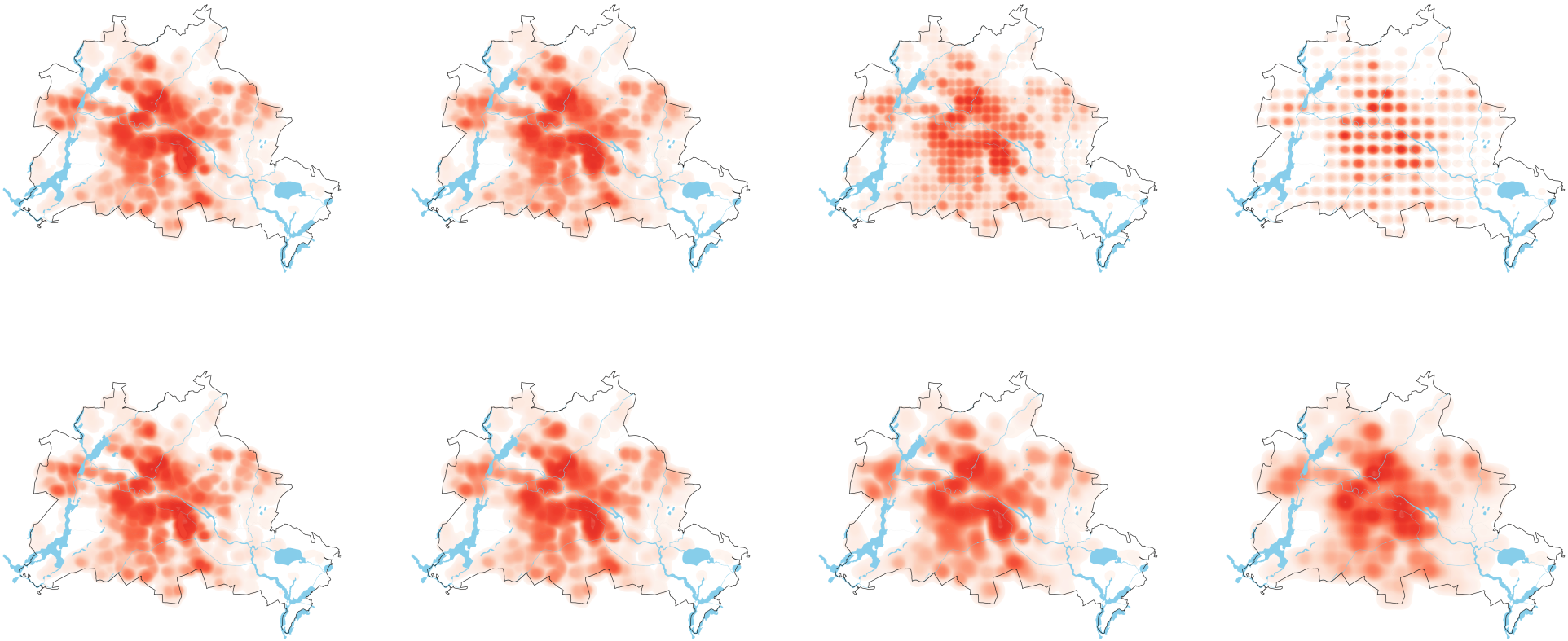
Application: Data Sources

- The data 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 errors.
- Registration at the local residents' office is compulsory in Germany and is carried out by the federal state authorities.
- One of the scenarios we explore is **rounding by using grids of size 2000 meters by 2000 meters** that approximately correspond to the LOR geography.
- The original data includes the total number of residents at their principal residence and the number of persons according to **some key demographic characteristics**:
  - Ethnic background (Ethnic)
  - Age (Age over 60).

Freie Universität Berlin

Density of population: Ethnic minority background

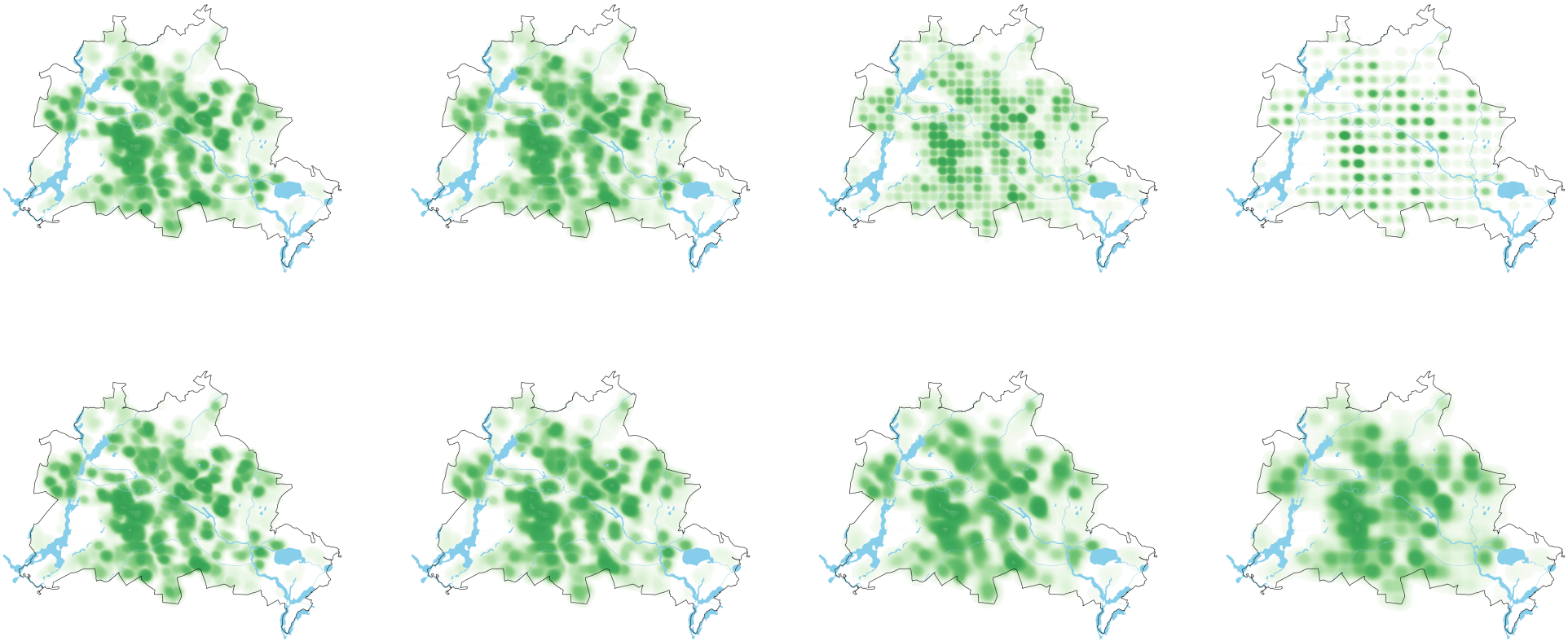
Naive (top panel) and MCMC estimators (bottom panel) with rounding step sizes of 0 (left), 500, 1250 and 2000 m (right).



Freie Universität Berlin

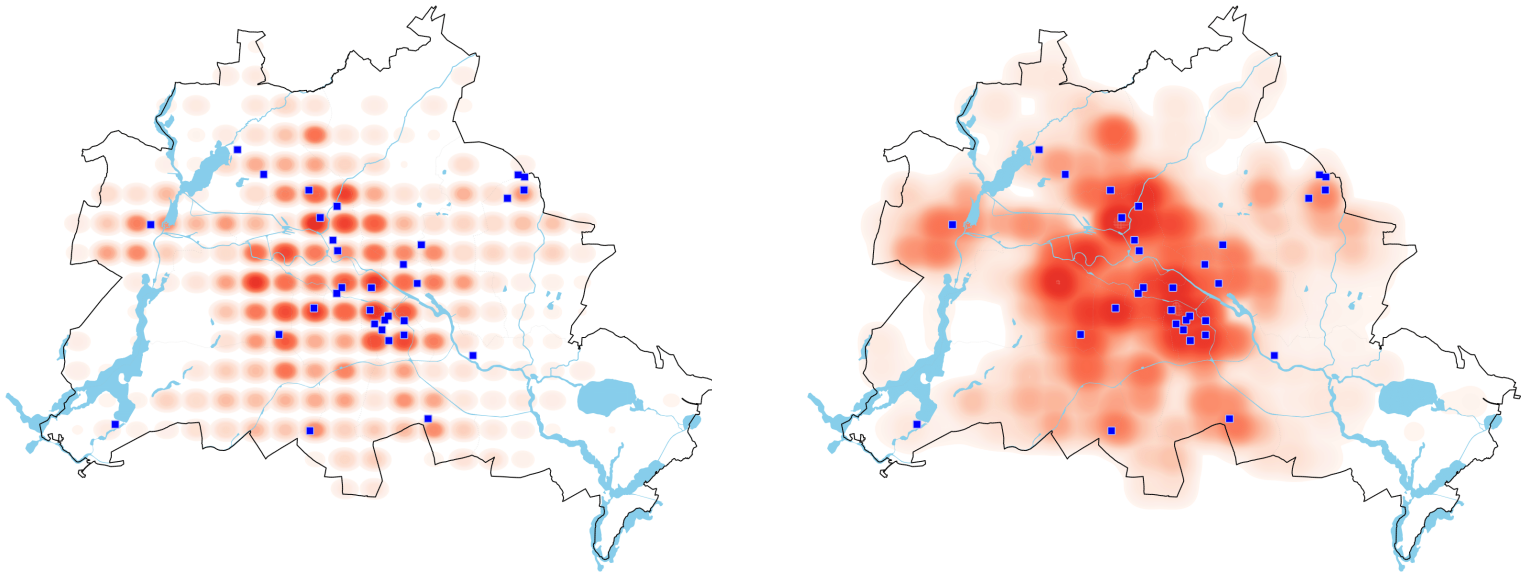
Density of population: Aged 60 and above

Naive (top panel) and MCMC estimators (bottom panel) with rounding step sizes of 0 (left), 500, 1250 and 2000 m (right).



Freie Universität Berlin

Advisory services for ethnic minorities

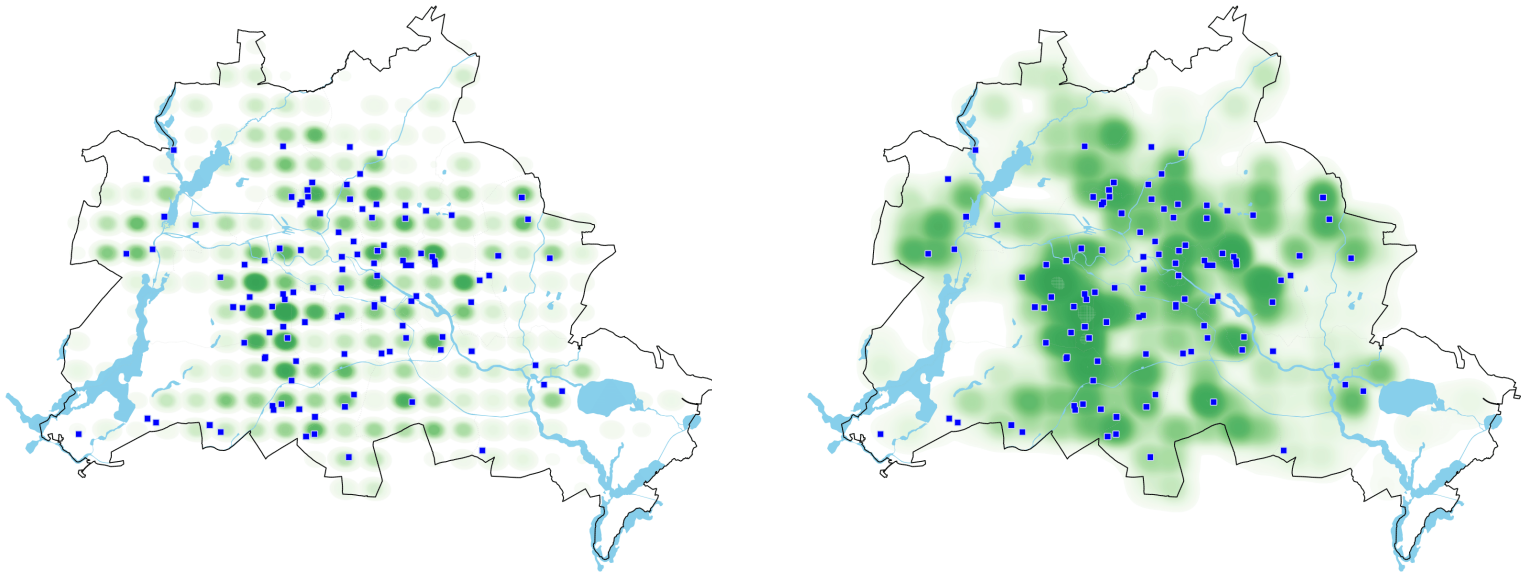


Ethnic background for rounding step size of 2000 m. Blue points indicate migrant advisory centres in Berlin.

- **Spatial distribution** of advisory centres **covers** ethnic minorities in the centre and north of Berlin quite well.
- **Some hotspots** with a high density of ethnic minority population but without any advisory service centres.

Freie Universität Berlin

Care for the elderly



Age above 60 for rounding step size of 2000 m. Blue points indicate retirement houses in Berlin.

- The **supply of retirement houses is good** in the center of Berlin.
- **Locations for future expansion** of retirement houses and other support structures can be identified, like some areas in the north (Reineckendorf) or in the south-east (Gropiusstadt).

Summary

- The proposed MCMC method **can offer considerably deeper insights**, compared to a Naive estimator, to data analysts about the density of target populations.
- The **structure preserving property** of the proposed MCMC method is particularly attractive when working with data that has been **subjected to disclosure control** via aggregation or rounding of the geographic coordinates.
- The use of the proposed methodology is facilitated by the **availability of a computationally efficient algorithm**.
- **Local authorities may prefer** the density estimates produced by the proposed estimator for making informed decisions.

Freie Universität Berlin

Thank you very much for your attention.

Ulrich Rendtel (Ulrich.Rendtel@fu-berlin.de)
Marcus Groß
Timo Schmid
Sebastian Schmon
Nikos Tzavidis