Department of Economics
Research Area: Statistics and Econometrics

Freie Universität Berlin

Estimating the density of ethnic minorities in Berlin

Multivariate kernel density estimation applied to sensitive geo-referenced administrative data protected via measurement error

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** in Figure 1.
- **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?

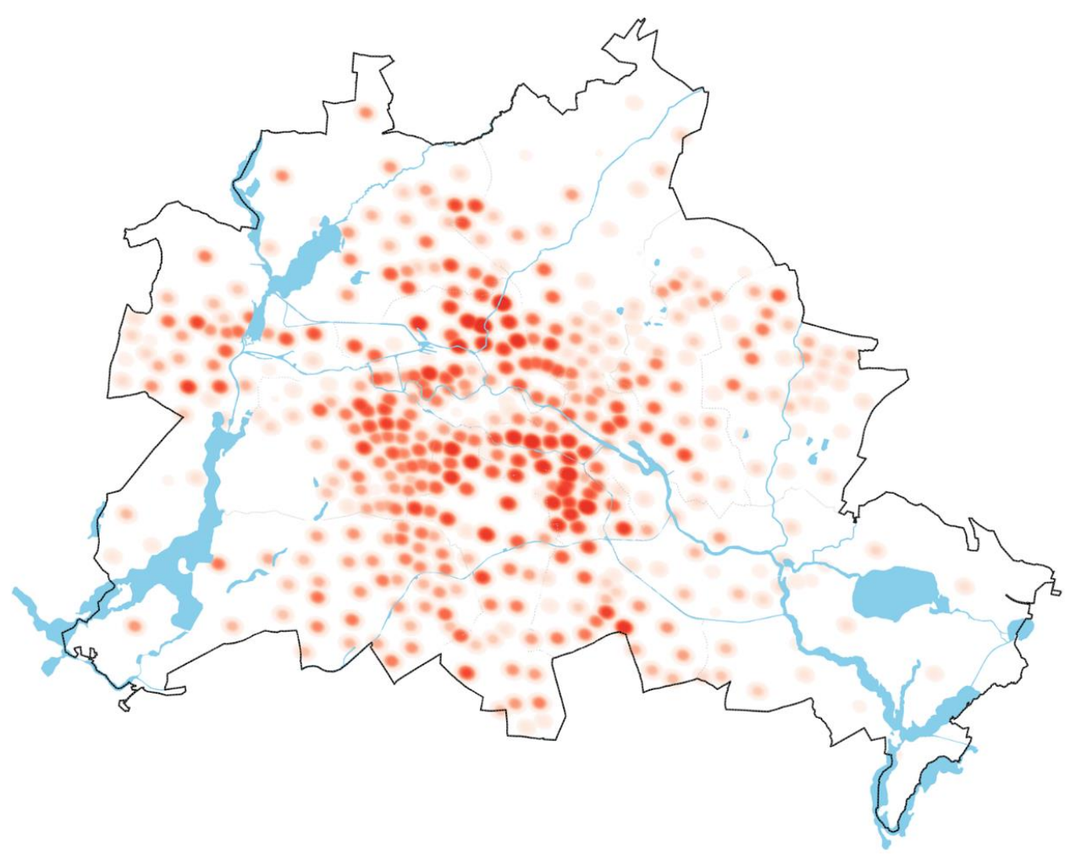

Figure 1: Density of the population of ethnic minority background in Berlin based on the publicly available data.

Methodology

- **Multivariate kernel density approach**: For a sample of bivariate data $X = \{X_1, \dots, X_n\}$ from a density f the kernel density estimate is defined as

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

- $K(\cdot)$ is a two dimensional kernel function and H is a bandwidth matrix.
- Through the anonymisation process only the **rounded values $\boldsymbol{W_i}$ are available**. This can be regarded as a measurement error on the true values X_i (Berkson, 1950).
- We formulate a **hierarchical Bayesian measurement error model**:

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

- The latent true values X_i are treated as additional parameters.
- Estimation is performed by an **iterative MCMC-type algorithm** which repeatedly draws synthetic samples of X_i from the square of side length r (rounding value) around W_i according to the current density estimate $\hat{f}(x)$.
- For details see Groß, M., Rendtel, U., Schmid, T., Schmon, S. and Tzavidis, N. (2015).

Berlin Register Data

- 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.
- 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.
- Around **950,000 people of ethnic background** from around 190 countries live in the 12 districts in Berlin.
- The three largest communities consist of approximately 200,000 people of Turkish ethnic background, around 100,000 people from Russia or from the former Soviet Union and approximately 60,000 people of ethnic background from the former Yugoslavia.
- The average number of residents of ethnic background at a household address is 3.07 with a median of 0.

Analysis of the Berlin Register of Residents

- The **benefits of using the proposed MCMC** estimator that accounts for measurement errors are illustrated using the Berlin register data.
- The application aims at estimating the **density of the ethnic minority population** in Berlin.
- We **impose grids** on the geographical space of the Berlin data set with respective grid sizes of 0 (original data), 500, 1250 and 2000 meters in Figure 2.
- The scenario by using grids of size 2000 meters by 2000 meters approximately corresponds to the LOR geography.
- We note that the **proposed MCMC** estimator (right panel) **outperforms** the Naive estimator (left panel) especially for large grid sizes.
- For grid sizes larger or equal to 1250m the Naive estimator produces small spikes at the location of the grid points.
- The proposed **MCMC estimator preserves the fundamental structure** of the underlying density.
- For the largest grid size (2000m), which implies strongly anonymised data, the general shape produced with the proposed MCMC estimator is clearly visible.

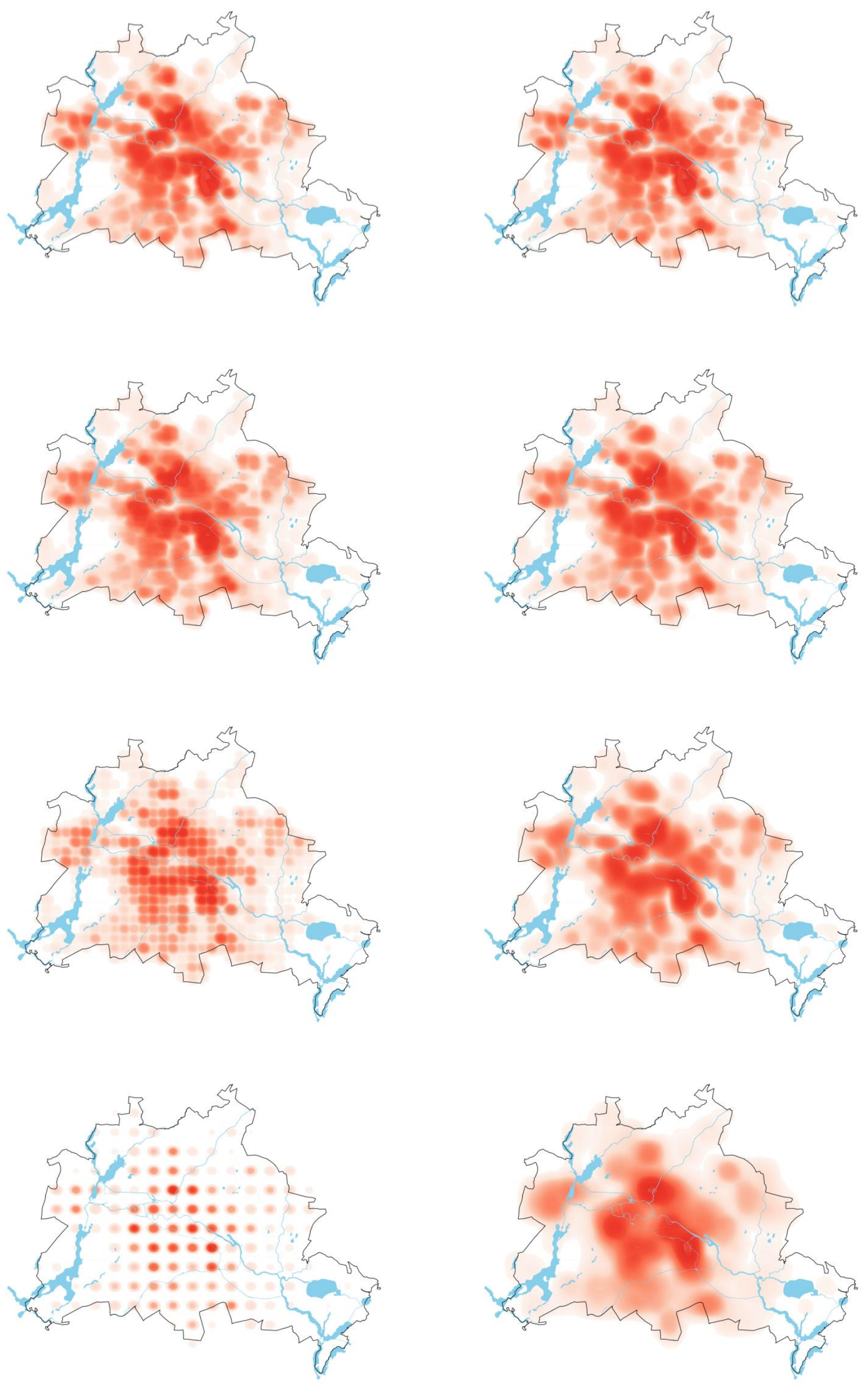

Figure 2: Naive (left panel) and MCMC estimators (right panel) with rounding step sizes of 0, 500, 1250 and 2000 m (top down).

Discussion and Summary

Discussion:

- The density of ethnic population is particular **high in the former West-Berlin districts** like Wedding, Neukoelln and Kreuzberg.
- The former German Democratic Republic Berlin districts (Friedrichshain and Prenzlauer Berg) show lower density.
- The **spatial distribution of advisory centres** covers ethnic minority populations in the centre and north of Berlin quite well.
- There are **some hotspots** with a high density of ethnic minority population but without any advisory service centres.

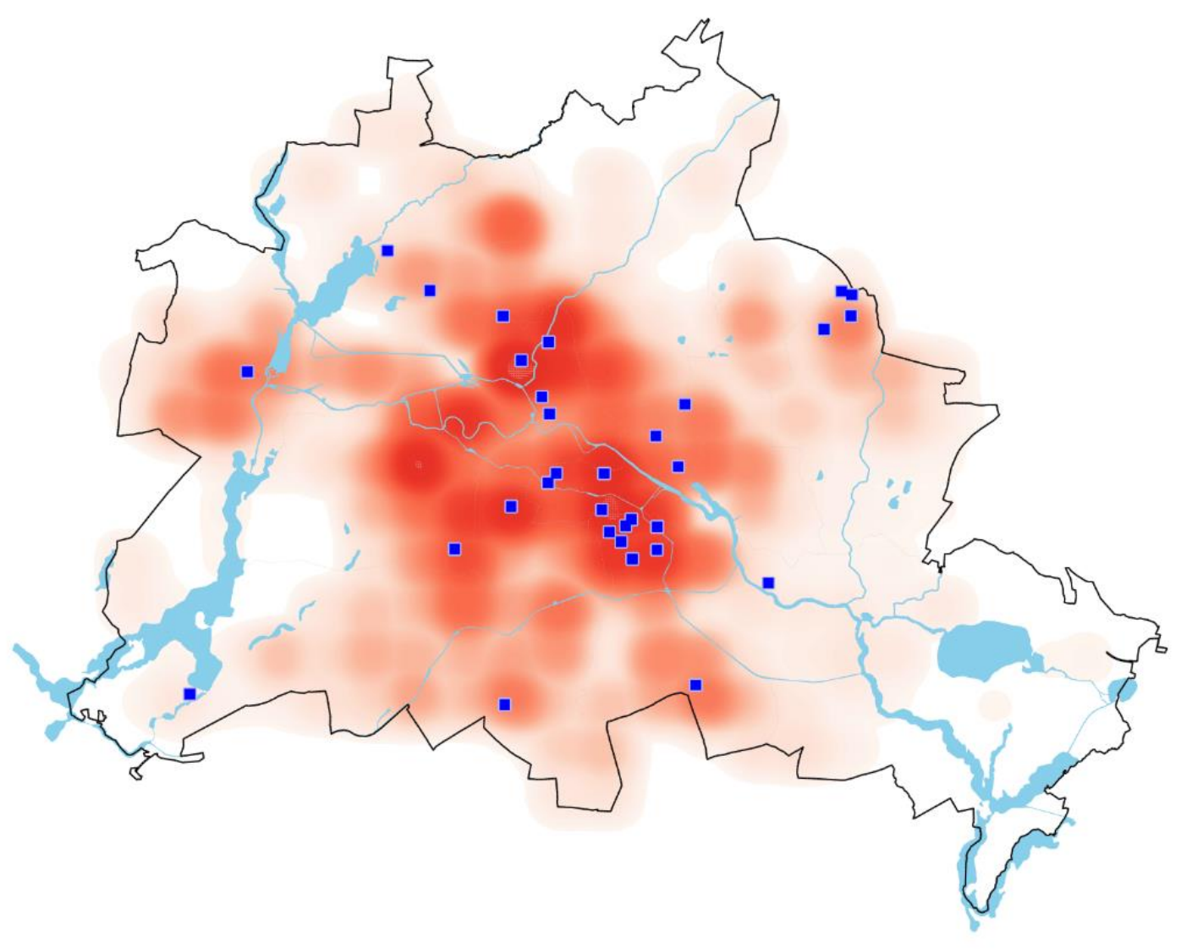

Figure 3: Ethnic background for rounding step size of 2000 m. Blue points indicate migrant advisory centres in Berlin.

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**.

References:

- Berkson, J. (1950): *Are there two regressions?,* Journal of the American Statistical Association, 45, 164-180.
- Groß, M., Rendtel, U., Schmid, T., Schmon, S. and Tzavidis, N. (2015): *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*, Discussion Paper 2015/7, School of Business & Economics, Freie Universität Berlin.

M. Groß*, U. Rendtel*, T. Schmid*, S. Schmon* and N. Tzavidis°

** Freie Universität Berlin* *° University of Southampton*