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 geocoded 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 (LDR) - 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?

**Methodology**

- Multivariate kernel density approach: For a sample of bivariate data \(X = (X_1, ..., X_n)\) from a density \(f\) the kernel density estimate is defined as

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

  \(K(\cdot)\) is a two-dimensional kernel function and \(H\) is a bandwidth matrix.
- Through the anonymisation process only the rounded values \(W_i\) 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:

  \[
  \pi(X, H|W) = \pi(W|X) \times \pi(X|H) \times \pi(H|Prior)
  \]

  \(\pi(X, H|W)\) is 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 \(H\) (rounding value) around \(W_i\) according to the current density estimate \(f(\cdot)\).

**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 3,468,619 (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 1.
- The scenario by using grids of size 2000 meters by 2000 meters approximately corresponds to the LDR 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.

**Discussion and Summary**

**Discussion:**

- The density of ethnic population is particular high in the former West-Berlin districts like Wedding, Neukolln 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.

**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:**


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

* Freie Universität Berlin  
* University of Southampton