Application and quality assessment of simulated geo-coordinates for regional analysis of the parliamentary elections for the Bundestag 2017

Keywords: *Kernel density estimation, Simulated geo-coordinates, Choropleth maps.*

1 Introduction

Data collection and data interpretation have become main issues in modern information society. Concerning this matter preparation and visualisation of data are essential aspects for any analysis. With an appropriate illustration one will get easy access to the underlying information of data. Therefore, the chosen illustration method can have a major influence on the interpretation of the data. For data containing geographic references, map presentations are frequently used as they are the best method to illustrate spatial relations with the help of coloured symbols, borders and areas. The data discretization and the colour codes for the categorisations may result in huge differences of its visual impact. When dealing with aggregated data the approach of pre-processing becomes a key issue for obtaining informative maps. To get comprehensive information for all geo-coordinates of the region of interest a new non-parametric approach for density estimation named "kernelheaping" [1] is applied and evaluated. Based on the election results for the German Bundestag in 2017 the new technique is compared against standard choropleth maps and some modifications, like normalisation. The iterative kernelheaping procedure is also compared to non-iterative variants of kernel density estimators [2]. In addition, the new approach is evaluated statistically with an underlying known real-world density on different aggregation levels.
The Master thesis was written in cooperation of the Statistical Office of Berlin-Brandenburg and the statistical department of the economic faculty of Freie Universität Berlin. Beforehand the author completed an internship at the office. In this internship the kernelheaping method was set up and tested on preliminary election data and subsequently put into practice on the election night by using real time data. The results of the regional analysis were published in the official election report and the journal of the Statistical Office of Berlin-Brandenburg.

2 Methods

The central method which is used here is the kernelheaping algorithm which iteratively generates a kernel density based on regional aggregates. This algorithm is an application of a Stochastic Expectation Maximization (SEM) algorithm which is derived from the classical EM algorithm. Here the E-Step is replaced by a draw from the current density. The sample of these simulated geo-coordinates is realized by stratified sampling where the strata sizes are given by the local aggregates. After some iterations the algorithm converges to a density with estimates of geo-coordinates for all units. For a successful application, a fixed number of iterations is needed. In

the analysis this parameter is systematically modified to point out differences in the outcomes and provide optimal parameter setups.
In addition, classical choropleth maps are applied on the same datasets. These are the methodological standard for the visualisation of aggregated datasets. For choropleth maps there is no need for exact geo-coordinates, but the main drawback is the homogeneous colour within regional units and the discontinuities at the borderlines of the units. Usually the colour categories are limited to five levels, which represents a major loss of information. Furthermore, regional units usually don't have the same size. Therefore, the interpretation of choropleth maps via the area sizes, which is the most appealing one, can lead to misleading conclusions. In the examples it is shown that simple area normalizations already help to create more realistic map illustrations.
For comparisons a naive kernel density estimation (KDE) approach is applied as well. It is a common smoothing technique which can be used to estimate a density for aggregated data. It assumes the units to lay in the middle of the reference area. One drawback is the complex task of finding suitable smoothing parameters (bandwidth). It is shown that even with optimally chosen parameters (based on a true density) the naive kernel density is not able to create results as good as the kernelheaping method.
All three methods are applied to get the cartographical presentation of the election results and are used for the comparison analysis based on a known density. For all methods systematic parameter adjustments were implemented to point out optimal parameter setups and to evaluate the techniques regarding their robustness. Since the true density of election data is usually unknown, and to accomplish a fair comparison, local data of eligible voters is used in the comparison analysis instead of voting data. Thanks to working in cooperation with the Statistical Office Berlin-Brandenburg it was possible to get access to an anonymised dataset of eligible voters for every address in Berlin. Hence, a true density of eligible voters could be generated to allow comparisons based on quantitative criteria with real-world data. The mean squared error (MSE) criterion is chosen as the most appropriate. On the one hand, an MSE can be computed for every pixel in every map (represented as an "MSE map"), and on the other hand, it is possible to calculate an overall mean over all pixels of an MSE map for every applied method. This is done for eight different levels of local aggregation which leads to additional conclusions about the extent of aggregation the used procedures provide reasonable results.

3 Results

As seen in Figure 1 maps based on the kernelheaping algorithm (left panel) optically outperform a choropleth map (right panel). While the kernelheaping map clearly displays areas with a high density of voters the choropleth map does not show-up any regional structure. To get to a density, the choropleth map is normalized. This just influences the values of the scale, but not the distribution of colours itself. Since voting districts are generated such that they cover approximately the same number of voters the uniform distribution of voters indicated by the choropleth map is a natural consequence. While the choropleth approach uses only one value per area the kernel heaping approach uses one value per pixel which is much more flexible and produces more realistic results.
The kernelheaping is outperforming pixel-based naive kernel density estimations (KDE) as well, even if an optimal KDE bandwidth is calculated based on the true density, which in practice is not available (not displayed in Figure 1). By using plug-in bandwidth selectors, the result gets even worse. The KDE produces some peaks at the centre of the regional districts. Especially when the districts are on the fringe of the city the spots are clearly visible. The new kernelheaping algorithm provides smooth

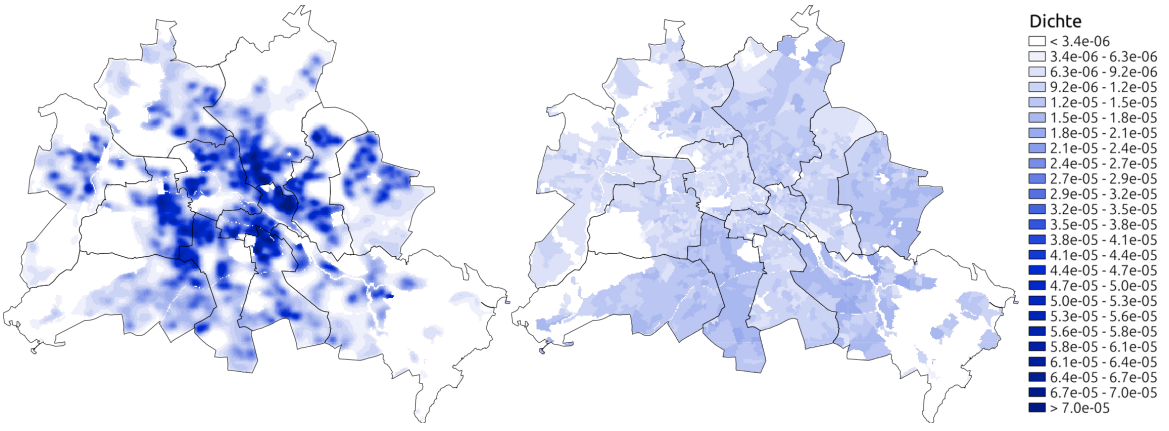

Figure 1: Density estimates of the distribution of eligible voters in Berlin with masked uninhabited areas. On the left panel calculations are done with the kernelheaping algorithm. On the right panel the choropleth map is applied on the same dataset.

transitions for larger districts as well as sharp contours for small districts in the centre of Berlin.
The local performance of all procedures is evaluated by biases, variances and resulting mean squared errors. The overall performance for all methods is described by the mean MSE over all pixels, resulting in an assessment criterion. Besides the already shown optical benefits, the kernelheaping procedure also outperforms all other methods during a quantitative analysis, as displayed in Figure 2. It appears that the performance for every method also depends on the level of aggregation, but among the applied methods, kernelheaping constantly provides the lowest MSE values. Only the standard choropleth maps (without area normalization) fail to achieve more accurate results using a lower level of aggregation. Note that the KDE with optimal chosen bandwidth based on the true density ("KDE Naive Optimal MSE") usually cannot be calculated in practice. On the lowest ballot box with aggregation level (UWB) the kernelheaping method gains about one half of the KDE MSE.

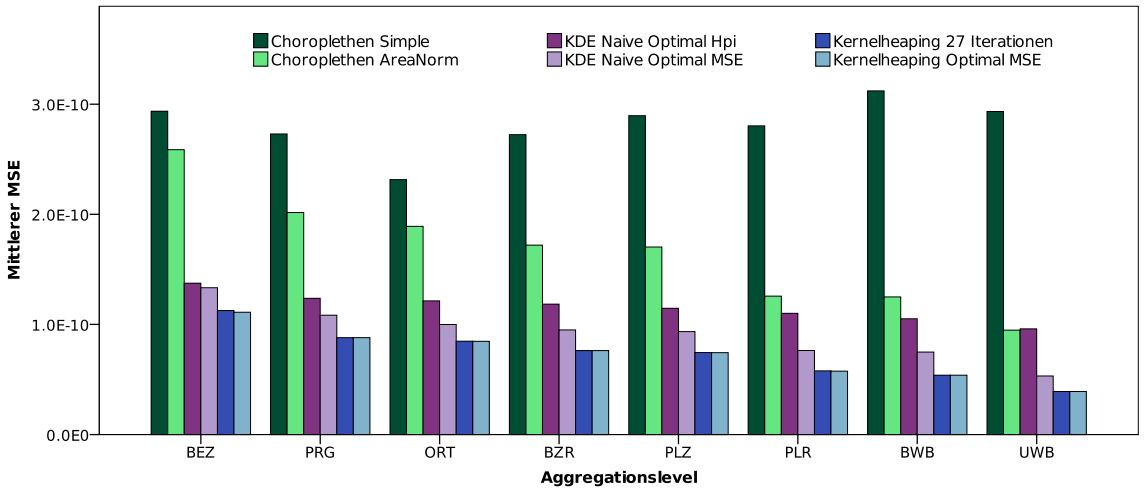

Figure 2: Mean MSE over all pixels depending on the level of aggregation and used methods. From left to right the aggregation level gets more detailed. Except for simple choropleth maps (dark green) the density estimation becomes better regarding the MSE for all methods applied. Furthermore, within each level of aggregation the kernelheaping results in the most accurate density estimation.

In addition to the MSE maps, analogous variance maps can be calculated for the kernelheaping procedure. As it is an iterative approach which depends on stratified

random sampling the method delivers slightly different results for each run. An interesting observation is the behaviour of the local variances at the centres and at the edges of the voting districts, as seen in Figure 3. It turns out that the variance is increasing when several district borders meet. Due to the iterative algorithm pixels close to the border are more influenced by the aggregated result of neighbouring districts and therefore have higher volatilities.

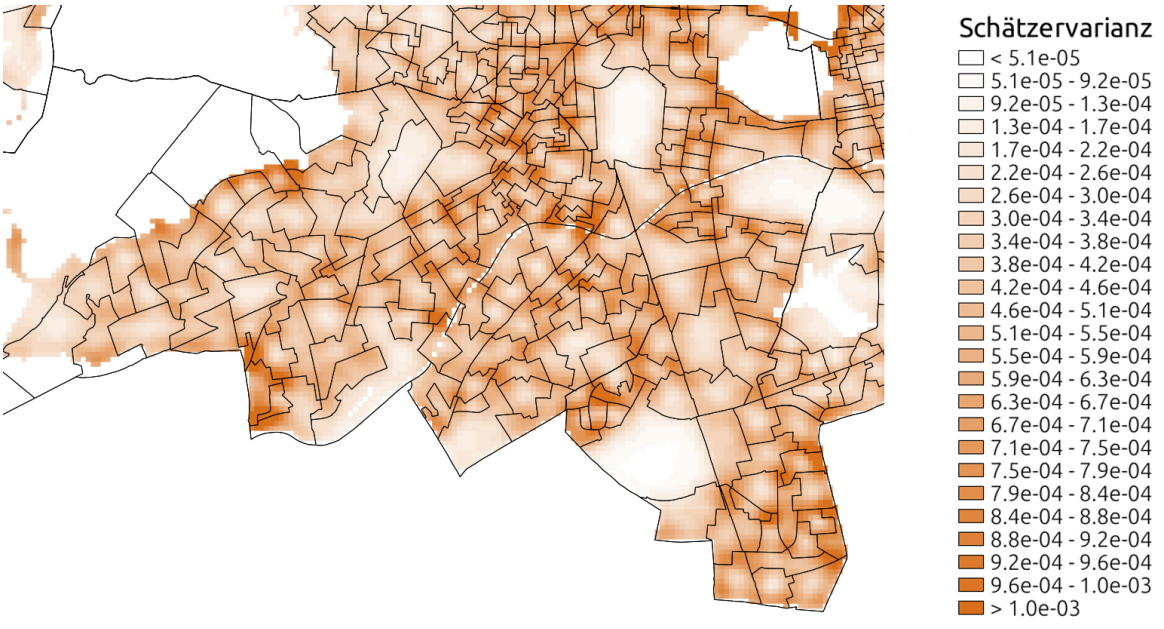

Figure 3: Variance for each pixel of the kernelheaping procedure over all iterations and Markov-chains. Close to the edges of the voting district the variance of the estimation becomes larger. The more districts hit each other the larger the variance appears to be.

Besides the preparation of maps, which were used for the official election report and the journal of the Statistical Office of Berlin-Brandenburg, the kernelheaping procedure has been optimized by the author. Some adjustments were made to speed up the runtime and allow an efficient analysis at the election night. Here the R-program was modified to parallel running independent Markov-Chains instead of one long running Markov-Chain only. This leads to more robust estimation results and makes it easier to parallelize the algorithm on multi-core processors. This reduces computation time as well.

4 Conclusions

A new statistical method was applied on a key area in official statistics: the presentation of regional voting results. Therefore, the potential of the kernelheaping approach was revealed. It was shown that it outperforms also naive standard density estimation tools when dealing with aggregated data. Since the kernelheaping procedure is not yet integrated as a standard tool in statistical offices, the thesis encourages the usage of this new technique. Furthermore, the Master thesis itself has become a decent documentation for employees to familiarize oneself with the procedure and incorporate it into further fields of application. Hence, the Statistical Office is equipped with a new statistical method from now on.
Practical improvements and modifications of the algorithm, which contribute to a lower runtime and more robust results, were implemented and the code of the kernelheaping package has been updated in R [3]. For the visualization with QGIS scripts

have been developed to automatically create cartographical maps with the obtained densities.

References

[1] Marcus Groß. *Messfehlermodelle für die Survey-Statistik und die Wirtschaftsarchäologie*. PhD thesis, Freie Universität Berlin, 2016.

[2] B.W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1986. ISBN 9780412246203.

[3] Marcus Groß. Kernelheaping: Kernel density estimation for heaped data., 2017. URL https://cran.r-project.org/web/packages/Kernelheaping/index.html. R package Version 2.0.