sciendo

Journal of Official Statistics, Vol. 36, No. 2, 2020, pp. 297–314, http://dx.doi.org/10.2478/JOS-2020-0016

Switching Between Different Non-Hierachical Administrative Areas via Simulated Geo-Coordinates: A Case Study for Student Residents in Berlin

Marcus Groß[1], *Ann-Kristin Kreutzmann*[1], *Ulrich Rendtel*[1], *Timo Schmid*[1], *and Nikos Tzavidis*[2]

The transformation of area aggregates between non-hierarchical area systems (administrative areas) is a standard problem in official statistics. For this problem, we present a proposal which is based on kernel density estimates. The approach applies a modification of a stochastic expectation maximization algorithm, which was proposed in the literature for the transformation of totals on rectangular areas to kernel density estimates. As a by-product of the routine, one obtains simulated geo-coordinates for each unit. With the help of these geo-coordinates, it is possible to calculate case numbers for any area system of interest. The proposed method is evaluated in a design-based simulation based on a close-to-reality, simulated data set with known exact geo-coordinates. In the empirical part, the method is applied to student resident figures from Berlin, Germany. These are known only at the level of ZIP codes, but they are needed for smaller administrative planning districts. Results for (a) student concentration areas and (b) temporal changes in the student residential areas between 2005 and 2015 are presented and discussed.

Key words: Choropleth maps; kernel density estimation; statistical reporting; sub-regional estimation; urban development.

1. Introduction

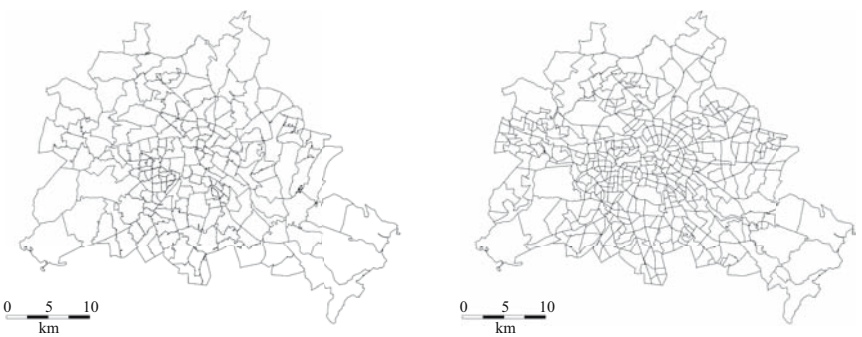
Maps are increasingly used for the dissemination of official statistics. Mostly, these consist of areas that display some value of interest in different colors. Thus, maps demonstrate, for instance, where the poor live (U.S. Census Bureau 2017), where people are most exposed to air pollution (Spiekermann and Wegener 2003), and where accessibility to services is low (Langford et al. 2008; Schmid et al. 2017). Therefore, maps are also an illustrative and easily understandable basis for targeting policies.

The commonly used maps are "choropleths" that use a discretization of the value of interest. The areas or zones are defined, for example, by administrative districts at different levels or statistical units as the European nomenclature des unités territoriales statistiques (NUTS), see the Statistical Atlas of the European Statistical Yearbook (Eurostat 2018). In using choropleth maps, it is problematic that the size of an area is not properly taken into account, which may lead to misinterpretations. Alternatively, areas can be defined by a

[1] Freie Universität Berlin, Garystraße 21, 14195 Berlin, Germany. Emails: Marcus.Gross@inwt-statistics.de, Ann-Kristin.Kreutzmann@fu-berlin.de, Ulrich.Rendtel@fu-berlin.de, and Timo.Schmid@fu-berlin.de
[2] University of Southampton, Murray Building 58, Highfield Campus, Southampton, UK. Email: N.Tzavidis@soton.ac.uk

rectangular grid of a certain size, say 1 km^2. These maps are often referred to as grid maps, see for an example the German Census atlas (Statistische Ämter des Bundes und der Länder 2015). Gallego et al. (2011) discuss several approaches that can be used to downscale area data to fine-scale raster grids in order to receive maps with a higher resolution and thus precision. Grid or raster data is commonly used in urban planning and simulations (Schürmann et al. 2002; Lautso et al. 2004; Patterson et al. 2011). Geo-coded data enables to create a different type of map that is independent of area definitions. These maps are based on a two dimensional kernel density of the variable of interest. They display each level of the estimated density by a different color. In contrast to choropleths, the color scheme, often ranging from light for low values to dark for high values of the density, is continuous. An example is the Service Map of Helsinki (OpenStreetMap Foundation 2019), where the user can combine different background maps with kernel density estimates of demographic subpopulations, like age groups and ethnic minorities.

In addition to the described downscaling or disaggregation of data to subsets of administrative units, switching between different area definitions/systems is often a challenge in official statistics. Performing statistical inference on an area level without available data while having data for another related area level is also known as spatial change of support (COS) (see e.g., Bradley et al. 2016). This occurs when there are different local planning areas in use, for example, fire brigade districts, schooling districts, hospital districts that are different from the standard administrative areas. For certain large-scale planning projects, such as an airport, the number of inhabitants in the upcoming noisefield of airplanes is of interest. European data in the INSPIRE Knowledge Base (Infrastructure for spatial information in Europe) are often reported on squares of different length. Here it may be necessary to adapt the European units to the local units (European Commission 2019). In the application of this work, the number of student residents in administrative areas of Berlin, "Lebensweltlich orientierte Räume" (LOR), is required by the Berlin Senate Department for Urban Planning and Environment for planning purposes. LORs are the smallest urban planning units for Berlin and have an average area size of around 1.99 km^2. However, the university enrollment registers only provide student totals at the level of ZIP codes with an average area size of around 4.62 km^2. Figure 1 shows the 447 LORs, as well as the 193 ZIP-code areas of Berlin. A careful inspection of the areas reveals many cross-cuttings of the area borders. Figure 2



(a) (b)

Fig. 1. ZIP-code areas of Berlin (a) and administrative planning areas (LORs) (b).

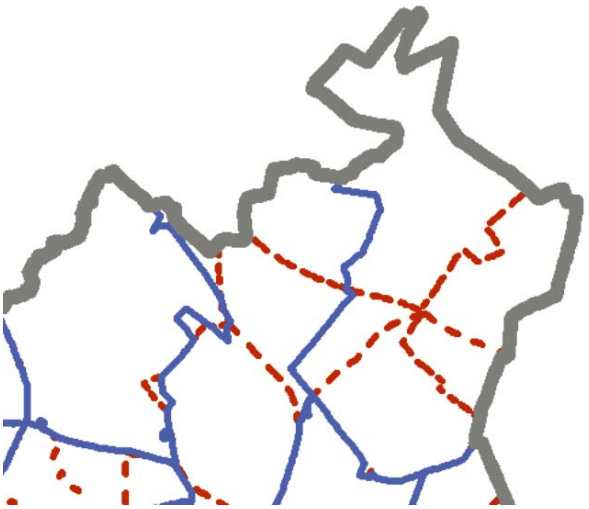


Fig. 2. The non-hierarchical structure of the ZIP areas (blue straight lines) and the LOR areas (red dotted lines) in the north east edge of Berlin.

demonstrates, in detail, the non-hierarchical structure of the ZIP-code areas (blue lines) and the LOR areas (dotted red lines) in the north east edge of Berlin. In other words, LORs are by no means a lower-level area system than ZIP-code areas.

As these two area systems are non-hierarchical, we are confronted with a problem that is hard to solve at an elementary level. Often this task is advanced by ad hoc methods, based on a proportional allocation of totals depending on which part of the ZIP-code area belongs to the respective administrative planning area (LOR). Such an approach is tedious and relies on an unrealistic assumption, namely, that the units are uniformly distributed across the ZIP-code area. Instead, Mugglin and Carlin (1998) and Mugglin et al. (1999) propose a hierarchical Bayesian approach for the spatial change of support. Bradley et al. (2016) extend the approach to data with sampling variability and enable the spatial and temporal change of support. These model-based approaches require covariate information on the area of interest and rely on distributional assumptions. Within the field of small area estimation, Trevisani and Gelfand (2013) extend hierarchical Bayesian models to soften the requirements for the covariate information by allowing the use of covariates of areas non-nested within the small areas of interest. In this work, we suggest a non-parametric alternative in the form of a kernel density estimate (KDE) that tackles the problem of transferring count numbers from one area system to another without covariate information. As the density function is independent of administrative areas, it is possible to compute count numbers for any area definition/system from the density.

In our case, we do not have the exact geo-coordinates at hand but only totals for areas that are not related to the areas of interest. Therefore, we present an approach in which geo-coordinates are simulated from area-specific aggregates. The method proposed in this work is similar to the approach of Groß et al. (2017), where it is used to counteract the rounding of geo-coordinates due to confidentiality reasons. In their analysis, kernel densities are generated to detect concentration areas of migrants and elderly persons in

Berlin. Rendtel and Ruhanen (2018) use the approach with "Open Data" in order to demonstrate local need for child care.

The algorithm of Groß et al. (2017) works for totals on rectangles that are the outcome of the rounding process. However, the approach can be extended to totals of arbitrary shape files. The algorithm is based on two elementary steps: the first step is to draw a sample from a two-dimensional density that gives the simulated geo-coordinates. The sampling is done with respect to the known number of observations in the reference areas, which is achieved by stratified sampling. The second step is a classical estimation step that generates a kernel density estimate from a sample of geo-coded data. These two steps resemble a "stochastic expecation maximization" (SEM) algorithm (Celeux et al. 1996).

The article is organized as follows. In Section 2, the proposed algorithm and its statistical foundation is described in more detail. Section 3 evaluates the quality of the conversion to different areas via a design-based simulation study. The proposed method is applied to the Berlin student residents data set in Section 4. Furthermore, the problem of allocating the students of Berlin to administrative areas/LORs for the planning of student homes and other student-related infrastructure is discussed. Besides the estimation of the total number of students in administrative areas/LORs, the kernel densities offer alternative methods to display regions with a dense student population and their development over time. The method is compared with the classical approaches via choropleths. Section 5 concludes and provides further research ideas.

2. Method

Multivariate kernel density estimation is a non-parametric approach to estimate the joint probability distribution of two or more continuous variables. Let $\boldsymbol{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ denote the exact geo-coordinates, such as longitude and latitude, of observations $i = 1,.., n$ with $\boldsymbol{X}_i = (X_{i1}, X_{i2})$. To estimate the density $f(x)$ at point x, a multivariate kernel density estimator is employed, which is given by:

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

where $K(\cdot)$ denotes a multivariate kernel function and $|\boldsymbol{H}|$ denotes the determinant of a symmetric positive definite bandwidth matrix $\boldsymbol{H}$. A popular choice for $K(\cdot)$ is the multivariate Gaussian kernel. The choice of $\boldsymbol{H}$ is highly important for the performance of the kernel density estimator. In principle, all bandwidth selection strategies try to minimize the mean integrated squared error (MISE) which is $E\int\left(\hat{f}_H(x) - f(x)\right)^2 dx$, where $f(x)$ is the true density and $\hat{f}_H(x)$ is the kernel estimate using the bandwidth matrix $\boldsymbol{H}$. For high case numbers, the asymptotic MISE (AMISE) offers some simplification by omitting some terms of lower order. The essential part depends on the mixed derivatives of the underlying density $\int f^{(m)}(x) f(x) dx$. Wand and Jones (1994) suggest some simple but efficient approximations of the empirical substitute $1/n\sum_{i=1}^{n} \hat{f}_H^{(m)}(x_i)$. They discuss the choice of the bandwidth in the multivariate case by using a plug-in estimator, which is also used in this work.

Instead of the exact geo-coordinates $\boldsymbol{X}$, only aggregated data for certain areas is available in this work. Simply applying a kernel density estimator to, for instance the area centers, leads to strongly biased estimates for rectangular shapes as shown in Groß et al. (2017). Therefore, a special treatment is needed. Following Groß et al. (2017), we can interpret the available data on area level, denoted by $\boldsymbol{W} = \{\boldsymbol{W}_1, \ldots, \boldsymbol{W}_n\}$, as a coarse measurement of the exact geo-coordinates of individual i. As the measurement process is known, we are able to formulate a measurement model $\pi(\boldsymbol{W}|\boldsymbol{X})$ for $\boldsymbol{W}$. It can be written as a simple product of Dirac distributions, $\pi(\boldsymbol{W}|\boldsymbol{X}) = \prod_{i=1}^{n} \pi(\boldsymbol{W}_i|\boldsymbol{X}_i)$, with

$$\pi(\boldsymbol{W}_i|\boldsymbol{X}_i) = \begin{cases} 1 & \text{for } \boldsymbol{X}_i \in area(\boldsymbol{W}_i), \\ 0 & \text{else}, \end{cases} \tag{2}$$

where $area(\boldsymbol{W}_i)$ stands for the set of geo-coordinates that belong to the area where $\boldsymbol{W}_i$ lies in.

From $\pi(\boldsymbol{X}_i|\boldsymbol{W}_i)$, we can draw pseudo samples of the $\boldsymbol{X}_i$ to estimate the density f by using the Bayes theorem:

$$\pi(\boldsymbol{X}_i|\boldsymbol{W}_i) \propto \pi\left(\boldsymbol{W}_i|\boldsymbol{X}_i\right)\pi(\boldsymbol{X}_i). \tag{3}$$

Thus, the exact geo-coordinates, $\boldsymbol{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$, are distributed according to the kernel density estimate restricted to the area where the observation $\boldsymbol{W}_i$ comes from. In an iterative procedure, the $\boldsymbol{X}_i$ are sampled from $\pi(\boldsymbol{X}_i|\boldsymbol{W}_i)$ followed by the estimation of $\pi(\boldsymbol{X}_i)$, respectively $f(x)$, by employing a multivariate kernel density estimator on the $\boldsymbol{X}_i$.

In particular, a SEM algorithm (Celeux et al. 1996) is utilized. This is a modification of the original EM-algorithm. The basic setting of the EM-algorithm refers to a situation where a part of the observations is missing. Thus, one has to maximize the marginal loglikelihood for the observed part of the data. As this can be quite complicated, one regards the expected value of the loglikelihood of the complete data where the expectation is done with respect to the current estimate of the parameter estimate. In many instances, this expected value of the loglikelihood of the complete data can be maximized by standard routines and leads to an update of the parameter estimate. With the SEM algorithm, the expected value is replaced by one realization of the unobserved part of the sample under the current parameter estimate. Again, the likelihood for this completed pseudo-sample is maximized and a new update of the parameter estimate is achieved. The generation of the pseudo-sample step brings a stochastic element into the algorithm, giving a more realistic distribution of the missing observations. In our application, the missing data are the exact geo-coordinates and the maximization of a likelihood is replaced by the kernel density estimation (a generalised SEM, Groß et al. 2017). In contrast to SEM, a simple EM algorithm would clearly not be helpful in this application, as all observations within an area would fall on the same location and thus not prevent a bias of the resulting kernel density estimate.

The algorithm starts with all the points concentrated at the center of the area. Starting from these artificial geo-coordinates, a kernel estimate $\hat{f}_{(0)}(x)$ is generated. Two iterative computation steps are performed as follows:

Step 1 (the 'S'-step in SEM): "Pseudo-samples" of the exact geo-coordinates, the $\boldsymbol{X}_i$, are drawn by sampling from the conditional distribution $\pi(\boldsymbol{X}_i|\boldsymbol{W}_i)$. This conditional

distribution is equal to the current density estimate restricted to the area where $\boldsymbol{W}_i$ belongs

Step 2 (the 'M'-step in SEM): The bivariate kernel density $\hat{f}_{(n+1)}(x)$ is estimated using the drawn pseudo-sample.

By $B + N$ iterations of Step 1 and Step 2, a sequence of kernel density estimates is generated. The final density estimate is computed by averaging the estimates of $\hat{f}_{(n)}(x)$ over the N samples after discarding the first B burn-in samples. The number of burn-in samples to achieve convergence and the number of samples N for a desired accuracy may depend on the application. As for MCMC-methods, no general recommendations can be given. However, for similar applications as presented here, we found that $B = 5$ and $N = 100$ was generally sufficient. More details on the kernel density estimation method and the exact implementation of the algorithm can be found in Groß et al. (2017). The only detail that is changed is to draw the pseudo-samples from the corresponding shape rather than from a rectangle, that means in the 'S'- step truncating the density to the area where observation $\boldsymbol{W}_i$ lies. This is more computationally intense, especially for complex-formed shapes, because we have to check whether there is a potential pseudo-sample inside the shape. However, this is of little importance with modern computers as long as the areas do not have a very complex shape, for example non-convex shapes cut into separate parts.

In our application, the problem arises from the fact that large areas of a town may consist of unsettled areas, like parks, lakes or industrial areas. These areas should be excluded from the generation of the geo-coordinates. If this information is available, it may considerably improve the estimation of the case numbers in the new area system. In principle, this is no problem for the SEM algorithm. One simply has to exempt the unsettled areas from the sampling of the geo-coordinates. However, in this case the boundary problem of the kernel density estimation comes into play, as the kernel function may not respect the boundary of the settled regions. One approach, the "cut- and normalize-method" (Gasser and Müller 1979), to overcome this problem is to restrict the kernel function to settled areas and to compute a new normalizing factor that makes the kernel function on the reduced area a density. Such a factor has to be computed for every spot where the kernel function is evaluated. This costs computational time, but it is not a real obstacle as it is implemented in existing software (Groß 2018).

After computing a non-parametric density estimate with this algorithm, the question arises how to allocate the number of observations to each shape in the new target area system. One possibility would be to numerically integrate over the non-parametric density and multiply the result by the number of total observations. However, it is likely that the result would not be consistent with the original data, that is, the number of observations belonging to a shape of the first area level would be different from the starting values. To preserve the original data structure, we chose to count the pseudo-samples falling in each shape of the target area system for each iteration. This also avoids numerical integration. These area counts will be averaged over all iterations.

The existence of N replications of an estimate makes it possible to calculate a confidence interval for the population value. In our simulation study below, we computed an interval that is given by the 5% and the 95% quantile of the N replications. This is not an exact confidence interval as it ignores the sampling from the density. But in our

application, sampling and its induced variance is not an issue as the starting values, that is, the numbers at ZIP-level, are population values with no variance. Nevertheless, the distribution over N replications reflects exactly the uncertainty of the knowledge of the case numbers in the new area system.

The algorithm is implemented in the R package *Kernelheaping* (Groß 2018) as function *dshapebivr*, which requires a data matrix with aggregated observation numbers for each area and a ∗.shp shapefile including the geometric data as input. The function *toOtherShape* in this package performs the operation to preserve the original data structure given the output of the function *dshapebivr* and an additional shapefile for the new administrative area system.

3. Simulation

In order to check the precision of the proposed method, we generated close-to-reality populations in a simulation. As a reminder, the areas of interest are the 447 LORs of Berlin, while the information of student totals is only given at the 193 ZIP-code areas. The cross-cutting of these area systems shown in Figure 3 confirms that the area systems are non-heirarchical.

We then randomly selected 15 mid-points to avoid a simple cluster in the center of the town. At each mid-point, 2,000 observations were generated from bivariate normal density with a variance of 3×10^6 (with covariance equal to 0). Then the points that were allocated to uninhabited areas were removed. Afterwards, we used two versions of the SEM algorithm. In the first version, we ignored the information about which areas are unsettled (SEM), while in the second version, we used the boundary correction (SEM-Boundary)

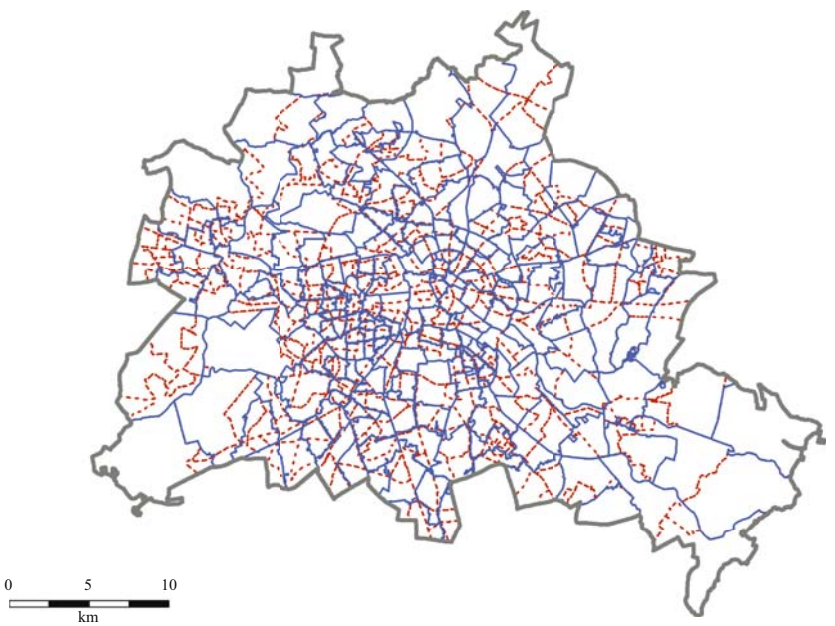


Fig. 3. Cross-cutting of ZIP-code area (blue, straight lines) and LOR area (red, dashed lines) borders in Berlin.

that keeps the density estimate within the settled areas. In order to evaluate the routine against the standard GIS procedure, we used a uniform density within the areas. Here, we used two versions as well. The first version ignores the unsettled areas (UNIFORM), while the second version respects the unsettled areas (UNIFORM-Boundary). The uniform allocation of the observations within the ZIP-code areas avoids the tedious computation of the cross-cutted areas that would be necessary in the GIS-approach, but is approximately equivalent. This procedure was repeated $R = 100$ times, however the preselected 15 mid-points were kept fixed. In order to compute 90-percent confidence regions, we selected the number of replications as $N = 100$. The burn-in phase was taken as $B = 5$.

Figure 4 displays one artificial allocation of geo-coordinates together with the LOR borders (left) and with the unsettled areas in green (forests and parks), blue (water) and grey (industrial and other).

In a next step, the number of observations falling in each area is counted at LOR-area level (treated as true values) and at the ZIP-code area level. The ZIP-code area level counts are used to estimate the "true" counts at the LOR-area level. As explained in Section 2, this is done by counting the number of the generated pseudo-samples falling in each LOR. There is no extra computational effort: during the generation of a new density, it can be checked in which of the LORs the new coordinates fall. Hence, every round of the SEM algorithm produces an estimate of the expected number of points falling into a LOR. Thus, it is only necessary to average these figures over the N Monte-Carlo replications.

Table 1 compares the performance of the four procedures with respect to the root mean squared error (RMSE) of the estimated LOR totals over the R replications, defined as

$$RMSE = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\widehat{LOR}_r - LOR_r\right)^2}$$

where $\widehat{LOR}$ denotes the estimated and LOR the true LOR total. The RMSE is computed for every area and the distribution of these area-specific RMSE values is then analyzed over areas. We see that the information on settled areas is helpful in reducing the RMSE

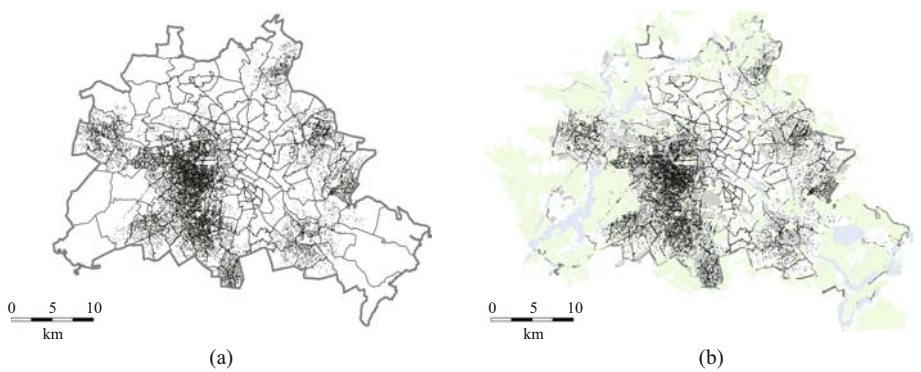


Fig. 4. Simulated geo-coordinates (a) and including their restriction to settled areas (b). Result of one out of 100 simulation runs.

Table 1. RMSE and coverage of the estimated LOR totals over the R = 100 simulation runs.

Method	Average RMSE	95% quantile RMSE	99% quantile RMSE	Max RMSE	Coverage of 90% quantile
SEM	5.63	9.91	24.85	45.4	83.54
UNIFORM	7.33	11.88	34.03	63.6	N/A
SEM-Boundary	4.36	7.45	15.93	38.70	90.54
UNIFORM-Boundary	5.04	9.19	16.24	47.05	N/A

for both algorithms. However, the average RMSE of the SEM algorithm is always lower than the average RMSE with the UNIFORM procedure. With no information on settled areas, the differences are more pronounced: here the SEM algorithm amounts to only 76.8% of the RMSE with the uniform distribution. With information on settled areas, the reduction drops to 86.6%. Similar figures are obtained for the upper quantiles of the RMSE.

The last column displays the coverage of the 90% quantile interval based on the replications of the SEM algorithm. For each area count there is such an interval. The area-specific coverage rates are then averaged over all areas. For our simulations, the average coverage of this interval is close to its nominal value. Thus, the variation of the N replications of the SEM algorithm may be used to construct a confidence interval for the area counts.

In order to demonstrate the use of the *Kernelheaping* package, we present a minimal working example in the Github repository Kernelheaping MinimalWorkingExample.

4. Application

The city of Berlin is a growing town. In the past five years, Berlin has gained around 220,000 people in total, see Senator für Stadtentwicklung und Umwelt (2016). A large proportion of this increase is due to the population gains in the age group of 20 to 30 years old, which contains many students. With the increasing number of students, questions for urban development planning arise. Where do students live and how do they get to their universities? What type of housing do students need? Students, as well as other social groups, have special requirements and behavioral patterns with regard to the local infrastructures.

To answer the above questions, it is helpful to have accurate and reliable information of the residential locations of students in Berlin. This information can improve the planning of projects that students benefit from and, consequently, these can be implemented more targeted. Therefore, the Senate Department for Urban Development and Environment aimed to analyze where students who are enrolled at Berlin universities are located in the metropolitan region of Berlin-Brandenburg and how they relate to the counts of LORs and Brandenburg municipalities. Before, there was no data available about student locations at small-scale residential areas. The LORs are the smallest urban planning units in Berlin. One possible data source about student residences are the enrollment offices of the Berlin universities. However, for privacy concerns, these figures are available only at the level of ZIP-code coordinates.

4.1. The Data

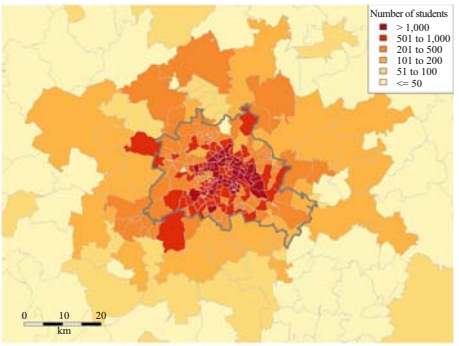
The number of students at ZIP-code area level in the years 2005, 2010 and 2015 could be established for the three – by far largest – universities of Berlin: Freie Universität Berlin (FU), Humboldt-Universität zu Berlin (HU) and Technische Universität Berlin (TU). The same applies for the rather small Alice Salomon Hochschule Berlin. Only for the year 2015, we were provided with numbers from Beuth Hochschule für Technik Berlin, the Hochschule für Wirtschaft und Recht Berlin (HWR) and the Hochschule für Technik und Wirtschaft Berlin (HTW). All numbers refer to the beginning of the winter term ('Wintersemester', abbr. WS), except for the data from FU and HU in 2015, which refer to the summer term ('Sommersemester', abbr. SoSe), where student numbers are typically lower. Thus, we applied a correction for the HU and the FU in 2015 and multiplied the numbers of these two universities by the ratio of winter term to summer term (e.g., FU: $36{,}674 = 33{,}173 \cdot 1.106$). Table 2 gives an overview of the total number of students in each year for every college and university, as well as the total number of students in Berlin. The data is provided by the Statistical Office for Berlin-Brandenburg. Figure 9 (see Appendix 6) visualizes the locations and size of the colleges and universities in Berlin. Furthermore, we have information on all dormitories in Berlin and the number of students living there for every considered year. As our information on ZIP-code totals does not cover all educational institutes with students in Berlin, our totals only sum up to 80% of the total number of students. With respect to the total number of students in Berlin, there is precise information from official statistical sources (Amt für Statistik Berlin-Brandenburg 2018). In order to cover the rest of the students from other institutes, we used some calibrations for the ZIP-code totals. As this calibration is not relevant for the method displayed here, we deferred the details of our calibrations to the appendix (see Appendix, Section 6). The students living in dormitories are not used for the kernel density estimates and are added afterwards to the final estimates at LOR-level to produce more accurate estimates, as their location is already known.

4.2. Results for the Location of Students in Different Map Representations

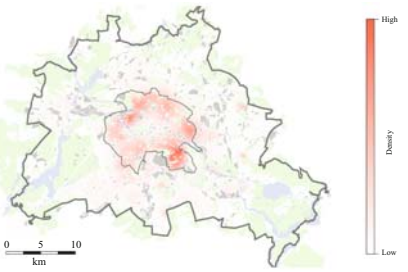
The maps in Figure 5 visualize the number of students in ZIP-code area (the level for which data is available), the kernel density estimate (transmission tool) computed on the

Table 2. Number of students in 2005, 2010 and 2015 for available colleges.

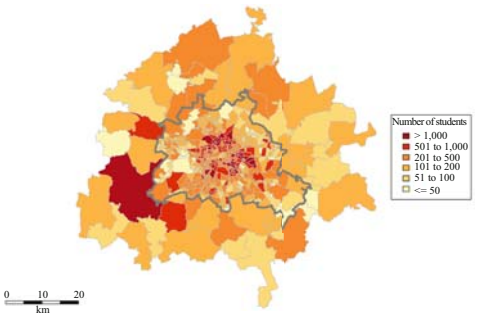
College/University	WS 2005	WS 2010	WS 2015	SoSe 2015
TU Berlin	29,772	29,758	33,933	-
FU Berlin	34,936	33,518	36,674	33,173
HU Berlin	32,428	29,689	34,214	31,098
Beuth	-	-	12,532	-
HTW	-	-	13,355	-
Alice Salomon	1,611	2,512	3,422	-
HWR	-	-	10,009	-
Sum of available colleges	98,697	95,477	144,139	-
Sum of all Berlin colleges	133,024	147,030	175,651	-



(a) Distribution at ZIP-code level



(b) Kernel density estimates



(c) Distribution at LOR level

Fig. 5. The plots show number (density) of Berlin students in 2015.

basis of these counts and the estimated number of students in the LORs of Berlin (the level of interest for urban planning) and its surrounding municipalities in 2015.

All three maps display a joint pattern with a concentration of students in a belt surrounding the center of the town. This belt is characterized by a traditional dense settlement (Senate Department for Urban Development and Housing 2017). It can also be seen that some students commute from neighboring municipalities to Berlin universities. Clearly, this number declines rapidly with the distance from Berlin. However, the graphical impression of the map with ZIP-code areas and LORs is quite different in the southwest of Berlin (the area of Potsdam). In the LOR representation, it looks very much as if there is a cluster that is densely populated with students. However, the ZIP-code area and the KDE representation do not exhibit such a pattern. The southwest "cluster" is simply the result of taking the entire municipality of Potsdam as one LOR.

When it comes to the individual development of the LORs with the highest student counts, it can be noted that they are located in certain districts of Berlin (Wedding, Neukölln, Moabit, Prenzlauer Berg, Friedrichshain and Kreuzberg). Table 3 lists the ten most popular LORs among students in 2015 and their development over time together with the 95% coverage interval for the 2015 values. The values exhibit remarkable changes in their student population from 2005 to 2015. Thus, the necessity of studies aiming to monitor the changes of the student population at a low level of regional aggregation is restated. With the exception of Rixdorf in Neukölln, all areas with a substantial increase of the student population lie in the north-west (Wedding and Moabit) of the central belt. In all the other LORs the student population is quite stable over time.

Table 3. The ten most popular urban planning areas (LOR) in 2015 with students counts for 2005, 2010 and 2015. The 95% coverage interval refers to the year 2015.

Urban planning area	District	2015	Coverage Interv.	2010	2005
Reuter Kiez	Neukölln	2072	(2051, 2094)	2187	1956
Samariterviertel	Friedrichshain	1892	(1833, 1940)	2095	2159
Rixdorf	Neukölln	1856	(1805, 1899)	1469	869
Rehberge	Wedding	1680	(1630, 1725)	1148	773
Traveplatz	Friedrichshain	1637	(1571, 1704)	1226	1354
Emdener Straße	Moabit	1580	(1566, 1591)	1162	942
Soldiner Straße	Wedding	1540	(1521, 1565)	1036	695
Reinickendorfer Straße	Wedding	1440	(1382, 1493)	923	603
Graefe Kiez	Kreuzberg	1409	(1393, 1426)	1350	1513

4.2.1. A Closer Look at the Temporal Development 2005–2015

Since data is available for the years 2005, 2010 and 2015, we can have a closer look at the temporal development of Berlin students residencies.

In general, the proportion of students living in Berlin has slightly but steadily increased from 82.3% in 2005 to 84.4% in 2015. In contrast to that, the percentage of students living in other German regions and foreign countries (mostly Poland) has decreased from 7.1% in 2005 to 5.0% in 2015. For a more detailed overview, Table 4 shows the estimated proportions of students living in Berlin, in the surrounding municipalities, in other municipalities of Brandenburg and outside of Berlin or Brandenburg. Focusing on Berlin and its surroundings, Figure 6 shows the KDE maps for each of the three reference years 2005, 2010 and 2015. From this representation, the structure of the students settlement seems to remain quite stable. However, if we display the highest density regions (HDR) remarkable regional changes can be seen. Note that such a representation is restricted to the KDE approach. Figure 7 compares the HDRs containing 25% and 50% of the students over time. Parts of the northwestern inner belt (Moabit and Wedding), as well as the southern belt (Neukölln) are now included in the 25% region in comparison to 2005. Parts of the eastern belt (southern Prenzlauer Berg and parts of Friedrichshain and Kreuzberg) did drop off from the 25% HDR in the last ten years. Interestingly, it becomes apparent that, in general, the concentration decreased. The 25% highest density region enfolded only 24.64 km^2 in 2005. This area grew to 28.58 km^2 in 2010 and to 33.27 km^2 in 2015. A

Table 4. Distribution of students of Berlin colleges living in Berlin, in the surrounding municipalities, in other municipalities of Brandenburg and out of Berlin/Brandenburg.

	2005	2010	2015
Berlin	109,436 (82.3%)	121,356 (82.5%)	148,231 (84.4%)
Surrounding municipalities	6,713 (5.0%)	7,648 (5.2%)	9,595 (5.5%)
Other municipalities of Brandenburg	7,504 (5.6%)	8,620 (5.9%)	9,059 (5.2%)
Other German regions and foreign countries	9,470 (7.1%)	9,406 (6.4%)	8,766 (5,0%)
Overall	133,024 (100%)	147,030 (100%)	175,651 (100%)

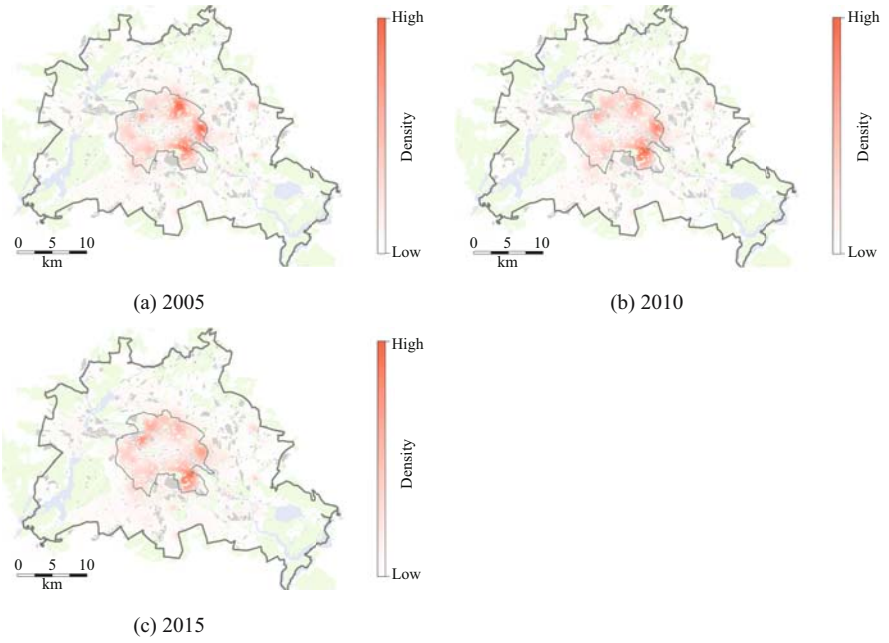


Fig. 6. The plots show the kernel density estimates of Berlin students in 2005 (a), 2010 (b) and 2015 (c).

similar effect is noticeable for the 50% HDR (2005: 76.88 km^2, 2010: 81.45 km^2, and 2015: 92.40 km^2).

Analyzing the absolute differences in the number of students on the level of the urban planning areas reveals further insights. Differences over the whole time period are visualized in Figure 8. A very large increase can be observed here for the locality of Wedding (northwest). The localities Neukölln (south), Lichtenberg (east), Moabit (northwest) and to a lesser extent Adlershof (southeast), Tempelhof (south) or Schöneberg (southwest) have gained students. Strong negative trends are recorded for Prenzlauer Berg (northeast) and the northern part of Mitte (center), which can be attributed to the

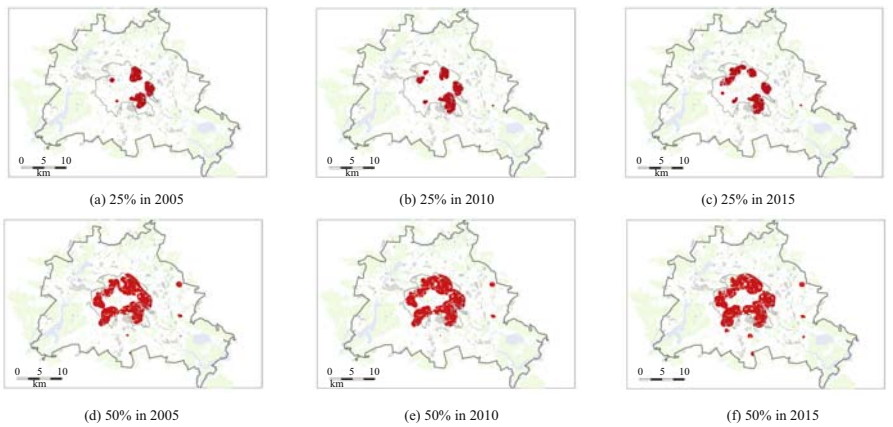


Fig. 7. Regions with highest student density: 25% of students (a, b, c) and 50% of students (d, e, f).

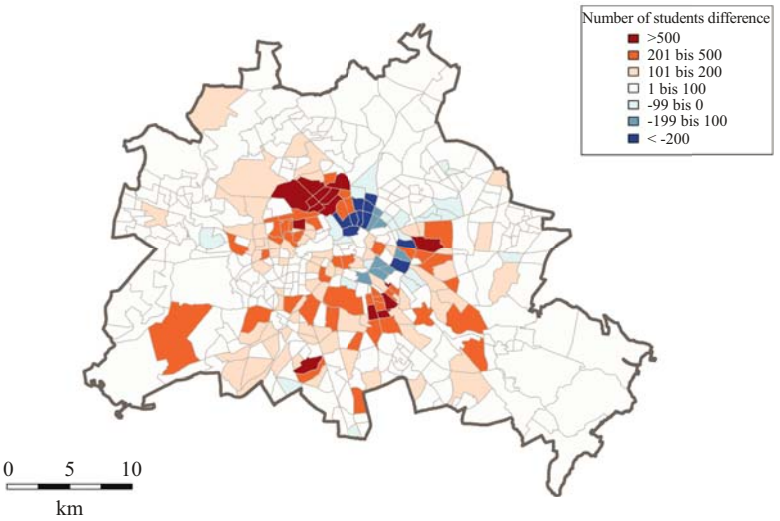


Fig. 8. Differences in student numbers 2015 compared to 2005 on administrative planning area level.

gentrification of these quarters (Schulz 2017; Holm and Schulz 2018). In addition, the eastern parts of Friedrichshain (east) and Kreuzberg (southeast) have lost students in the reference period.

The observations described may be due to the general increase of student numbers by almost 32% in Berlin, see Table 4. But they are also the result of a tightening housing market, which led the students to search for an apartment in other areas where housing is affordable for them. By contrast, Moabit, Wedding and Neukölln are propagated in the discussion on revaluation and displacement processes that can be carried out by pioneers such as students.

5. Conclusion

This work shows that kernel density estimates are a useful tool for the transformation of case numbers between area systems that are not hierarchical. Compared to ad hoc solutions, the proposed method is particularly preferable due to the following reasons. First, our approach is not based on the unrealistic assumption that the characteristic is uniformly distributed within areas. Second, while ad hoc solutions are often carried out manually, the approach in this work is available in the R package *Kernelheaping* and thus, the user can do this task quite automatically. Third, the algorithm is able to deal with uninhabited areas, which is a problem that is often encountered in practice. Fourth, the algorithm delivers coverage intervals for the population values. Finally, the proposed method is superior to the ad-hoc approach with respect to the RMSE.

Furthermore, density estimates, which are used as a transmission tool in this work, have their own merits. They help highlight the highest density regions, which can be used to identify local concentrations in the region of interest.

It should be noted that our algorithm is extremely useful for the construction of maps that are based on Open Data. Because of confidentiality reasons and their easy access, they

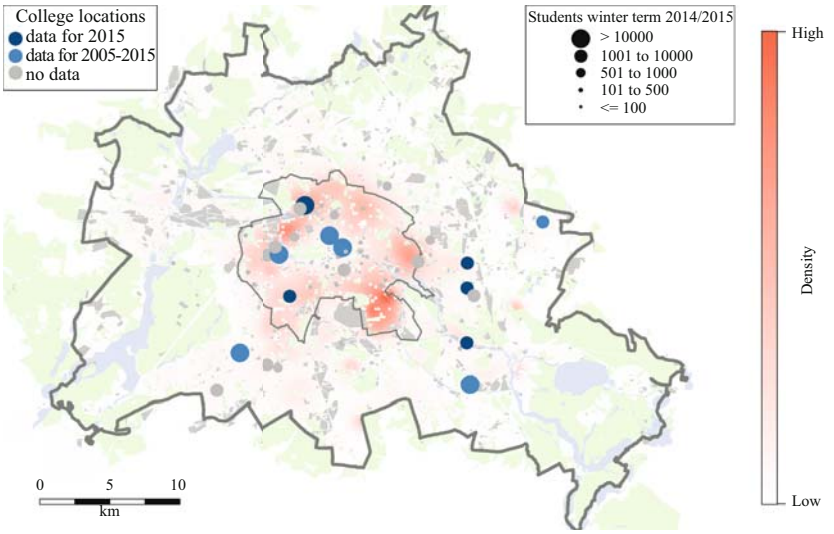


Fig. 9. Locations of colleges and universities of Berlin with number of students including the kernel density estimate of the student distribution in 2015. The border of the 'inner city' is added to the map.

are often provided as local aggregates. For example, the Open Data in Berlin are presented at the level of LORs or at a grid level (Berlin Open Data 2019). In the considered application the disclosure risk of individuals is not increased as the simulated geo-coordinates of individuals of a certain ZIP-code area are all drawn from the same distribution. However, additional information at individual level, such as ethnic affiliation, might help to identify an individual's location more precisely by running the presented algorithm on different sub-groups.

6. Appendix

The vast majority (about 80%) of Berlin's students in 2015 was covered by our sample of colleges and universities. Nevertheless, we would clearly underestimate the number of students in the planning areas due to the missing colleges. Therefore, a calibration is

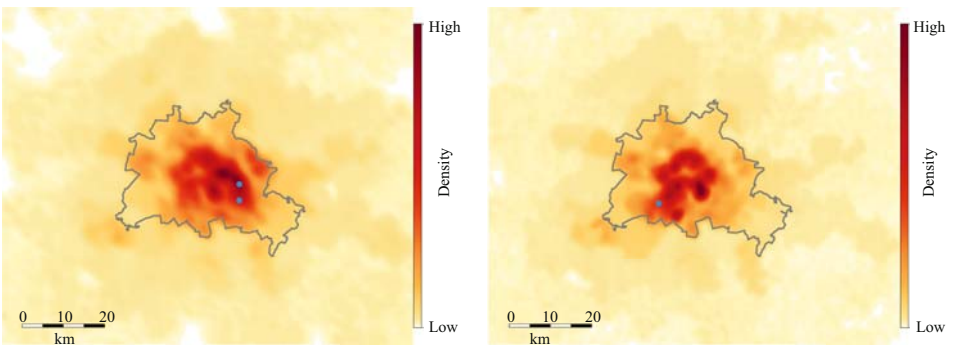


Fig. 10. Kernel density estimates of HTW (a) and FU (b) student distributions with college site locations.

necessary. The Statistical Office for Berlin-Brandenburg provides the total numbers of students enrolled in Berlin, giving us the possibility to simply upscale the total number of students in each ZIP-code area by a factor (e.g., multiplying by 175,651/144,139 = 1.22 for 2015; cf. Table 2). Another issue is the problematic comparison of the years 2005 and 2010 with 2015, as the coverage of colleges and universities is lower in these years. This is especially important as the specific college has a definite influence on the students' living address. We found out that a large proportion of the students live within the inner city borders, but some live near the college as well, as the kernel density estimate for 2015 shows (cf. Figure 9).

For the year 2015, we think that the effect of missing colleges is negligible, as we have information on the most important ones and the remaining ones are rather small and quite similarly distributed. If we would leave out the colleges only available in 2015, we get quite different area aggregates for ZIP-code areas near the missing colleges, for example ZIP code 10318 with only 145 instead of 796 students. Figure 10 excellently shows the kernel density estimates of the HTW and the FU student distributions. To account for the lower number of colleges in 2005 and 2010, we tried to adjust the number of students using the data of 2015. To achieve this, we employed a generalized linear mixed model, GLMM, (McCulloch and Neuhaus 2001) linking the number of students in each ZIP-code area considering all colleges available (Y) with the number considering colleges with data available for 2005 to 2015 (X). With a random intercept for each ZIP code ($zip_i \sim N(0, \tau)$), we fitted a Poisson-glmm with a log-link and the following model formula:

$$Y_i = \exp(\beta_0 + log(X_i + 1)\beta_1 + zip_i)$$

This formula was then used to predict Y for the years 2005 and 2010.

7. References

Amt für Statistik Berlin-Brandenburg. 2018. *Statistiken zu Bildung und Kultur*. Available at: https://www.statistik-berlin-brandenburg.de/BasisZeitreiheGrafik/Zeit-Hochschulen.asp?Ptyp=400&Sageb=21003&creg=BBB&anzwer=7 (accessed February 2019).

Berlin Open Data. 2019. *Datensätze*. Available at: https://daten.berlin.de/datensaetze (accessed February 2019).

Bradley, J.R., C.K. Wikle, and S.H. Holan. 2016. "Bayesian spatial change of support for count-valued survey data with publication to the American Community Survey." *Journal of the American Statistical Association* 111(514): 472–487. DOI: https://doi.org/10.1080/01621459.2015.1117471.

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. DOI: https://doi.org/10.1023/B:STCO.0000039481.32211.5a.

European Commission. 2019. *Inspire knowledge base - infrastructure for spatial information in Europe*. Available at: https://inspire.ec.europa.eu/ (accessed May 2019).

Eurostat. 2018. *Statistical Atlas*. Available at: http://ec.europa.eu/eurostat/statistical-atlas/gis/viewer/# (accessed January 2019).

Gallego, F.J., F. Batista, C. Rocha, and S. Mubareka. 2011. "Disaggregating population density of the European Union with CORINE land cover." *International Journal of Geographical Information Science* 25(12): 2051–2069. DOI: https://doi.org/10.1080/13658816.2011.583653.

Gasser, T. and H.-G. Müller. 1979. "Kernel estimation of regression functions." In *Smoothing techniques for curve estimation*: 23–68. Springer.

Groß, M. 2018. *Kernelheaping: Kernel Density Estimation for Heaped and Rounded Data*. R package version 2.2.0. Available at: https://cran.r-project.org/web/packages/Kernelheaping/ (accessed May 2020).

Groß, M., U. Rendtel, T. Schmid, S. Schmon, and N. Tzavidis. 2017. "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." *Journal of the Royal Statistical Society: Series A* 180(1): 161–183. DOI: https://doi.org/10.1111/rssa.12179.

Holm, A. and G. Schulz. 2018. GentriMap: "A model for measuring gentrification and displacement." In I. Helbrecht, ed. *Gentrification and Resistance*: 251–277. Wiesbaden: Springer VS.

Langford, M., G. Higgs, J. Radcliffe, and S. White. 2008. "Urban population distribution models and service accessibility estimation." *Computers, Environment and Urban Systems* 32(1): 66–80.

Lautso, K., K. Spiekermann, M. Wegener, I. Sheppard, P. Steadman, A. Martino, R. Domingo, and S. Gayda. 2004. Planning and research of policies land use and transport for increasing urban sustainability. Report, European Commission. Available at: http://www.spiekermann-wegener.de/pro/pdf/PROPOLIS_Final_Report.pdf (accessed February 2019).

McCulloch, C.E. and J.M. Neuhaus. 2001. *Generalized linear mixed models*. Wiley Online Library. DOI: https://doi.org/10.1002/9781118445112.stat07540.

Mugglin, A.S. and B.P. Carlin. 1998. "Hierarchical modeling in Geographic Information Systems: Population interpolation over incompatible zones." *Journal of Agricultural, Biological and Environmental Statistics* 3(2): 111–130. DOI: https://doi.org/10.2307/1400646.

Mugglin, A.S., B.P. Carlin, L. Zhu, and E. Colon. 1999. "Bayesian areal interpolation, estimation, and smoothing: An inferential approach for geographic information systems." *Environment and Planning* 31: 1337–1352. DOI: https://doi.org/10.1068/a311337.

OpenStreetMap Foundation. 2019. *Service Map*. Available at: https://servicemap.hel.fi/?municipality=helsinki&_rdr=Default.aspx (accessed January 2019).

Patterson, Z., M. Kryvobokov, F. Marchal, and M. Bierlaire. 2010. "Disaggregate models with aggregate Two UrbanSim applications." *Journal of Transport and Land Use* 3(2): 5–37. DOI: https://doi.org/10.5198/jtlu.v3i2.113.

Rendtel, U. and M. Ruhanen. 2018. "Die Konstruktion von Dienstleistungskarten mit Open Data am Beispiel des lokalen Bedarf an Kinderbetreuung in Berlin." *AStA Wirtschafts- und Sozialstatistisches Archiv* 12(3–4): 271–284. DOI: https://doi.org/10.1007/s11943-018-0235-y.

Schmid, T., F. Bruckschen, N. Salvati, and T. Zbiranski. 2017. "Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: estimating literacy rates in Senegal." *Journal of the Royal Statistical Society: Series A* 180(4): 1163–1190. DOI: https://doi.org/10.1111/rssa.12305.

Schulz, G. 2017. "Aufwertung und Verdrängung in Berlin – Räumliche Analysen zur Messung von Gentrifizierung." *WISTA – Wirtschaft und Statistik* 4: 287–314. Available at: https://www.destatis.de/DE/Methoden/WISTA-Wirtschaft-und-Statistik/2017/04/aufwertung-verdraengung-berlin-042017.html (accessed April 2020).

Schürmann, C., R. Moeckel, and M. Wegener. 2002. "Microsimulation of urban land use." ERSA conference papers, European Regional Science Association. August 27th–31st, 2002, Dortmund, Germany.

Senate Department for Urban Development and Housing. 2017. *06.06. population density*. 2017 edition. Available at: https://www.stadtentwicklung.berlin.de/umwelt/umweltatlas/edm606_04.htm (accessed January 2019).

Senator für Stadtentwicklung und Umwelt. 2016. *Bevölkerungsprognose für Berlin und die Bezirke 2015–2030*. Available at: https://www.stadtentwicklung.berlin.de/planen/bevoelkerungsprognose/download/2015-2030/Bericht_Bevprog2015-2030.pdf (accessed February 2019).

Spiekermann, K. and M. Wegener. 2003. "Modelling urban sustainability." *International Journal of Urban Sciences* 7(1): 47–64. DOI: https://doi.org/10.1080/12265934.2003.9693522.

Statistische Ämter des Bundes und der Länder. 2015. *Zensus Atlas*. Available at: https://atlas.zensus2011.de/ (accessed January 2019).

Trevisani, M. and A. Gelfand. 2013. "Spatial misalignment models for small area estimation: a simulation study." In N. Torelli, F. Pesarin, and A. Bar-Hen, eds. *Advances in Theoretical and Applied Statistics Studies in Theoretical and Applied Statistics*: 269–279. Berlin, Heidelberg: Springer.

U.S. Census Bureau. 2017. *Small area income and poverty estimates (saipe)*. Available at: https://www.census.gov/data-tools/demo/saipe/saipe.html?s_appName=saipe&map_yearSelector=2017&map_geoSelector=aa_c (accessed January 2019).

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

Received March 2019
Revised October 2019
Accepted December 2019