Freie Universität Berlin

Switching between different area systems via simulated geo-coordinates: A case study for student residents in Berlin

Marcus Groß
Ulrich Rendtel
Timo Schmid
Nikos Tzavidis

School of Business & Economics
Discussion Paper

Economics

2018/2

Switching between different area systems via simulated geo-coordinates: A case study for student residents in Berlin.

Marcus Groß[*], Ulrich Rendtel[†], Timo Schmid[‡], Nikos Tzavidis[§]

February 2018

Abstract

The transformation of area aggregates between non-hierarchical area systems is a standard problem of official statistics. We introduce a new method which is based on kernel density estimates. It is a modification of the SEM algorithm proposed by Gross et al. (2016), which was used for the transformation of totals on rectangular areas to kernel densities 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 a new area system.

The method is applied to student resident figures from Berlin. These are known only at the level of ZIP codes but they are needed for administrative planning districts. Our method is evaluated on a similar, simulated data set with known exact geo-coordinates. In the empirical part results for changes in the student residential areas between 2005 and 2015 are presented. It is demonstrated that the transformation via kernel density estimates offers additional useful features to display concentration areas.

Keywords: Choropleths, Grid Maps, Kernel Density Estimation, Geo-Coordinates

1 Introduction

Maps in official statistics are commonly created by areas that are displayed in different colors which display some value of interest. Usually, these so-called choropleths, use a discretization of the value of interest. The areas are defined by administrative districts at different levels, say NUTs 1, NUTs 2 or lower, see the Statistical Atlas of the European Statistical Yearbook (`http://ec.europa.eu/eurostat/statistical-atlas/gis/viewer/#`) However, different area systems may be in use that are not ordered in a hierarchical fashion. Alternatively, areas are defined by a rectangular grid of different size, say 1 km. These maps are often referred as grid maps, see for an example the German Census atlas (`https:`

[*] INWT Statistics

[†] Freie Universität Berlin

[‡] Freie Universität Berlin

[§] University of Southampton

//atlas.zensus2011.de/) With geo-coded data one would be able to create a different type of map that is independent from area definitions. These maps base on a two dimensional kernel density of the variable of interest. This style of maps displays each levels of the estimated density by a different color. Often a continuous color scheme is used ranging from light for low values to dark for high values of the density. An example is the Service Map of Helsinki (https://servicemap.hel.fi/?municipality=helsinki&_rdr=Default.aspx), where the user can combine different background maps with kernel density estimates of demographic subpopulations, like age groups and ethnic minorities.

The kernel density estimate can also help to tackle the problem to transfer count numbers from one area system to another. As the density function is not linked to areas it is possible to compute from the density count numbers for any area system. In the example treated here, the number of student residents in administrative areas of Berlin was of interest while the enrollment registers of the universities did only deliver student totals at the level of ZIP codes. As these two area-systems are non-hierarchical, one is confronted with a problem that is hard to solve at an elementary level. Often this task is advanced by ad-hoc methods which base on a proportional allocation of totals due to which part of the ZIP area belongs to the respective administrative area. Such an approach is tedious and relies on an unrealistic assumption, namely, the units are uniformly distributed across the ZIP area.

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. In this article we present an approach where we simulate geo-coordinates from area-specific aggregates. The method is similar to the the approach of Groß et al. (2016), who describe its use to counteract the rounding of geo-coordinates due to confidentiality reasons. In their analysis kernel densities were generated to detect concentration areas of migrants and elderly persons in Berlin.

The algorithm of Groß et al. (2016) works for totals on rectangles which are the outcome of the rounding process. However, their approach can be easily extended to totals of deliberate areas. The algorithm bases on two elementary steps: The first step is to draw a sample from a two dimensional density which 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 which generates a kernel density estimate from a sample of geo-coded data. These two steps resemble a so-called Stochastic EM (SEM) algorithm, see Celeux et al. (1996). The algorithm starts with all the points concentrated at the center of the area. Starting from this artificial geo-coordinates a new kernel estimate is generated. Then the sampling and estimation step is repeated in an iterative process. The algorithm does not only generate a final kernel density but also in each sampling step a set of simulated geo-coordinates. These geo-coordinates may be used to allocate the points to differently defined areas. Finally, these area counts will be averaged over the replications like the final density estimates. This feature circumvents the conputation of the volume under the density estimate for the new areas.

The article is organized as follows: In the methodological part we display the proposed algorithm and its statistical foundation in more detail. In the application part we consider the problem to allocate the Students of Berlin to small administrative areas for the planning of student homes and other student-related infrastructure. To get an assessment on the quality of the conversion to different areas, a simulation study is performed. The proposed method is then applied to the Berlin student residents. Besides the estimation of the total number of students in administrative areas the kernel densities offer alternative methods to display regions with a dense student population and their development over time. The methods are confronted with the classical approaches via choropleths.

2 Methods

Let $X = \{X_1, \ldots, X_n\}$ denote the exact geo-coordinates of the observations, with $X_i = (X_{i1}, X_{i2}),$ with $i = 1, .., n$. To estimate the density $f(x)$ at point x, a multivariate kernel density estimator is employed, which is given by:

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

$K(\cdot)$ denotes a multivariate kernel function. A popular choice is the multivariate gaussian kernel. H denotes a bandwidth matrix. The choice of H is highly important for the performance of the kernel density estimator. See e.g. Wand and Jones (1994), who discusses the choice of the bandwidth in the multivariate case by using a plug-in estimator, which is used here.

As we do not have the exact geo-coordinates but only aggregated data for certain areas, a special treatment is needed. This is because applying a kernel density estimator to e.g. the area centers leads to strongly biased estimates as shown in Groß et al. (2016) for rectangular shapes. Following Groß et al. (2016) we can interpret the available data on area level, denoted by $W = \{W_1, \ldots, W_n\}$, as data contaminated with measurement error. As the measurement error process is known we are able to formulate a measurement error model $\pi(W|X)$ for W. It can be written as a simple product of Dirac distributions, $\pi(W|X) = \prod_{i=1}^{n} \pi(W_i|X_i)$, with

$$\pi(W_i|X_i) = \begin{cases} 1 & \text{for } X_i \in Area(W_i) \\ 0 & \text{else.} \end{cases} \tag{2}$$

Now we can draw pseudo samples (imputations) of the X_i from $\pi(X_i|W_i)$, whereby the latter expression can be calculated by using the Bayes theorem:

$$\pi(X_i|W_i) \propto \pi(W_i|X_i)\pi(X_i) \tag{3}$$

Thus, the exact geo-coordinates, $X = \{X_1, \ldots, X_n\}$, are distributed according to the kernel density estimate restricted to the area where the observation W_i comes from. In an iterative procedure the X_i are sampled from $\pi(X_i|W_i)$ followed by estimation of $\pi(X_i)$, respectively $f(x)$, by employing a multivariate kernel density estimator on the X_i.

In particular, a Stochastic Expectation Maximization algorithm (SEM, Celeux et al. 1996) was utilized, as in Groß et al. (2016). The algorithm starts with all the points concentrated at the center of the area. Starting from this artificial geo-coordinates a new kernel estimate is generated. Two iterative computation steps are performed afterwards. The first step (the 'S'-step in SEM) is to draw so-called pseudo-samples of the exact geo-coordinates, the X_i, by sampling from the conditional distribution $\pi(X_i|W_i)$. This conditional distribution is equal to the current density estimate restricted to the area where W_i belongs to. In the second step (the 'M'-step in SEM), the bivariate kernel density $f(x)$ is estimated by using the drawn pseudo-samples. After a burn-in phase one may generate a sequence of kernel density estimates. The final density estimate is computed by averaging the estimate of $f(x)$ over all samples after discarding the burn-in samples. Details on the kernel density estimation method and the exact implementation of the algorithm can be found in Groß et al. (2016).

The only detail that needs to be changed is to draw the pseudo-samples from the corresponding shape instead from a rectangle, that means in the 'S'-step truncating the density to the area where observation W_i lies in. This is more computational intense, especially for complex formed shapes, because we have to check whether a potential pseudo-sample is inside the shape. However, this is of little importance with modern computers as long as the shapes do not have a very high complexity.

The algorithm is implemented in the R-package *Kernelheaping* (Groß, 2016) 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.

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 compliant with the original data, i.e. the number of observations belonging 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. These area counts will be averaged over all iterations. The function *toOtherShape* in the *Kernelheaping* package performs this operation given the output of the *dshapebivr* and an additional shapefile for the new area system.

3 Application for the allocation of students in Berlin

3.1 The setting of the study

The city of Berlin is a growing town. In the past five years Berlin has gained around 220,000 people in total. A large proportion of this 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 arouse. Where do students live and how do they get to their universities? Which type of housing do students need? Which infrastructures such as daycare centers, railway stations, bicycle parking facilities, bike paths, green spaces and cultural facilities are demanded and used? Students have, as well as other social groups, special requirements and behavioral patterns on facing the aforementioned infrastructures.

To answer the above questions, it is helpful to have accurate and reliable information of the residential locations of students in Berlin. Starting from the residential areas can improve the planning and implement student projects for their benefit more targeted[1]. So far, there are no data of student locations at small-

[1] The project 'Determination of the distribution of student accommodation' of the Senate Department for Urban Development and Environment in collaboration with the Freie Universität Berlin analysed 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, see the final report under `http://www.stadtentwicklung.berlin.de/planen/basisdaten_stadtentwicklung/studentisches_wohnen/index.shtml`.

scale residential areas. Here the so-called LORs[2] are the smallest urban planning units in Berlin. One possible data source on student residences are the enrollment offices of the Berlin universities. For privacy concerns these figures are available only the level of ZIP coordinates.

Figure 1 shows the 193 ZIP-code areas as well as the 447 LORs of Berlin. A careful inspection of the areas displays many cross-cuttings of the area borders, see Figure 2. So LORs are by no means a lower-level area system of ZIP-areas.

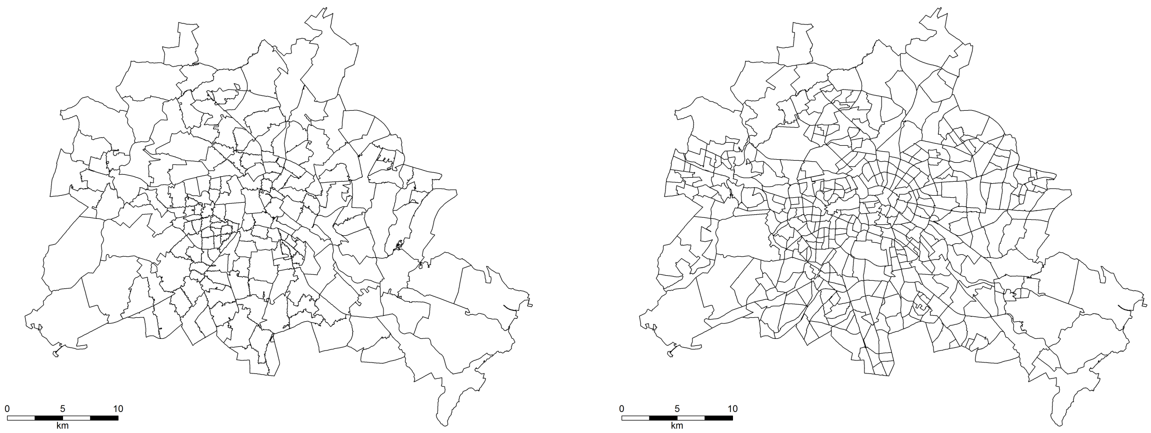

Figure 1: ZIP-code areas of Berlin (left) and administrative planning areas (LORs, right).

3.2 A simulation study with artificial data

In order to check the precision of the proposed routine, we generated hypothetical populations in a simulation. For each of the 447 LOR areas we generated $n_{LOR} = 250$ artificial geo-coordinates from an uncorrelated bivariate normal distribution with mean equal to the centroid of the LOR area and standard deviations of 1000 meters. As not all co-ordinates fall into the Berlin area there will be minor losses in observation counts. Totally, a sample of about 105000 observations is generated within each simulation, which is roughly comparable to the total number of students in our application example. Figure 3 displays one artificial allocation of geo-coordinates together with the LOR borders as well as the kernel density estimation based on these coordinates.

Now the number of observations falling in each coordinate is counted on LOR-area level and on the ZIP-code area level. The ZIP-code area level counts shall then be used to estimate the "true" counts on the LOR level afterwards. As explained in the methods section 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 an LOR. Thus it is only necessary to average there figures over the simulation runs. As a benchmark estimate we use the ad hoc approach of the introduction, which assumes that the observations are distributed uniformly inside the ZIP-code areas. This approach can be approximated by the SEM algorithm by replacing $\pi(X_i)$

[2] The acronym is derived from the German "Lebensweltlich orientierte Räume", which can be translated as life secular areas.

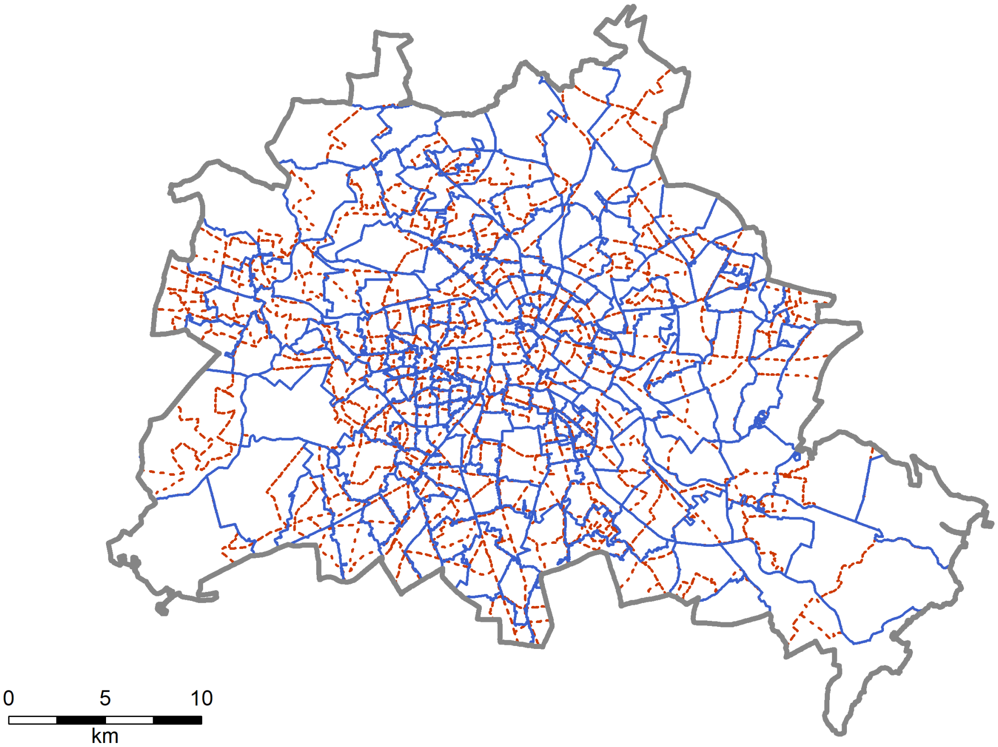

Figure 2: Cross-cutting of ZIP-code area (blue, straight lines) and LOR area (red, dashed lines) borders in Berlin.

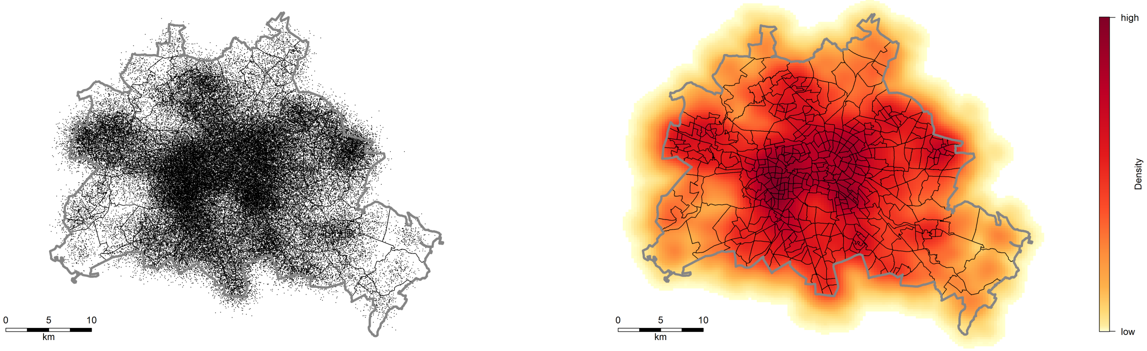

Figure 3: Simulated geo-coordinates (left) and Kernel density estimate based on simulated coordinates (right)

with the uniform distribution in equation (3). The whole procedure was repeated $n_{sim} = 100$ times with independent random draws for the geo-coordinates.

The overall structure of the LOR counts is well reflected by the SEM approach. Averaged over the $n_{sim} = 100$ simulations the mean absolute relative error over the 447 LORs is 9.8 percent for the SEM method and 14.4 percent for the ad hoc approach, while the RMSE of the departures of reference and the SEM totals amounts only to 33.5 persons compared to the RMSE of the ad-hoc method is 60.6 persons. Thus, in the considered simulation scenario, the presented SEM approach gives a considerable advantage over a simple approach assuming uniform distributions of observations within the shapes. Tables 1 and 2 present further details on the performance of both considered methods.

Table 1: Results of the simulation study: mean of RMSE measures over the $n_{sim} = 100$ simulation runs.

Method	Average RMSE	95% Quantile RMSE	99% Quantile RMSE	Max RMSE
SEM	33.5	62.2	114.8	317.3
AD HOC	60.6	102.1	201.3	730.0

Table 2: Results of the simulation study: mean of absolute percentage deviance (APD) measures over the $n_{sim} = 100$ simulation runs.

Method	Average APD	95% Quantile APD	99% Quantile APD	Max APD
SEM	9.8 %	29.7%	47.2 %	73.7%
AD HOC	14.4%	44.4%	68.8 %	85.8%

Finally, we want to compare whether there are some regional pattern in the absolute relative deviation of the SEM estimates as well as the ad hoc estimator and their reference values. By comparing Figure 4 with Figure 2 we find that (not surprisingly) the regional differences are bigger for areas with severe cross-cuttings of ZIP and LOR borders. Also for smaller areas the deviations seem generally smaller than for larger ones.

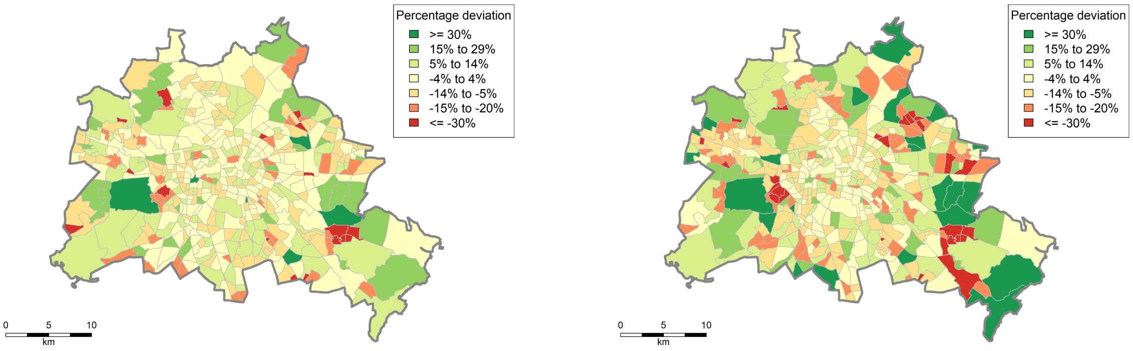

Figure 4: The relative deviation of SEM (left) and ad hoc method (right) compared to their reference values averaged over all simulation runs

3.3 The data

The number of students on 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 (FU), Humboldt Universität (HU) and Technische Universität (TU). The same applies for the rather small Alice Salomon Hochschule. Only for the year 2015 we were provided with numbers from Beuth Hochschule, the Hochschule für Wirtschaft und Recht (HWR) and the Hochschule für Technik und Wirtschaft (HTW). All numbers refer to the beginning of winter term ('Wintersemester', abbr. WS), except for the data of FU and HU in 2015, which refer to summer term ('Sommersemester', abbr. SoSe). Table 3 gives an overview on the total number of students in each year for every college and university as well as the total number of students in Berlin (data source: Statistical Office for Berlin-Brandenburg). Figure 11 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 there for every considered year.

Table 3: 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	-
Σ available colleges	98,697	95,477	144,139	-
Σ all Berlin colleges	133,024	147,030	175,651	-

As our information on ZIP totals covers not all Berlin educational institutes with students our totals sum up only to 80 percent of the total Berlin student numbers. With respect to the total number of students in Berlin there is precise information from official statistical sources. In order to cover the rest of the students from other institutes we used some calibrations for the ZIP totals. As this calibration is not relevant for the method displayed here we deferred the details of our calibrations to the appendix.

3.4 Results

Table 4 shows the estimated proportions of students living in Berlin, in the surrounding municipalities, in other municipalities of Brandenburg and out of Berlin or Brandenburg. 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 from foreign countries and other German regions as decreased from 7.1% in 2005 to 5.0% in 2015.

3.4.1 The location of students in 2015 in different map representations

The maps in Figures 5 to 7 visualize the absolute number of students in ZIP-Code area, the kernel density estimate computed on the basis of these counts and the estimated absolute number of students in the

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 municipalites	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%)

LORs of Berlin and its surrounding municipalities in 2015.

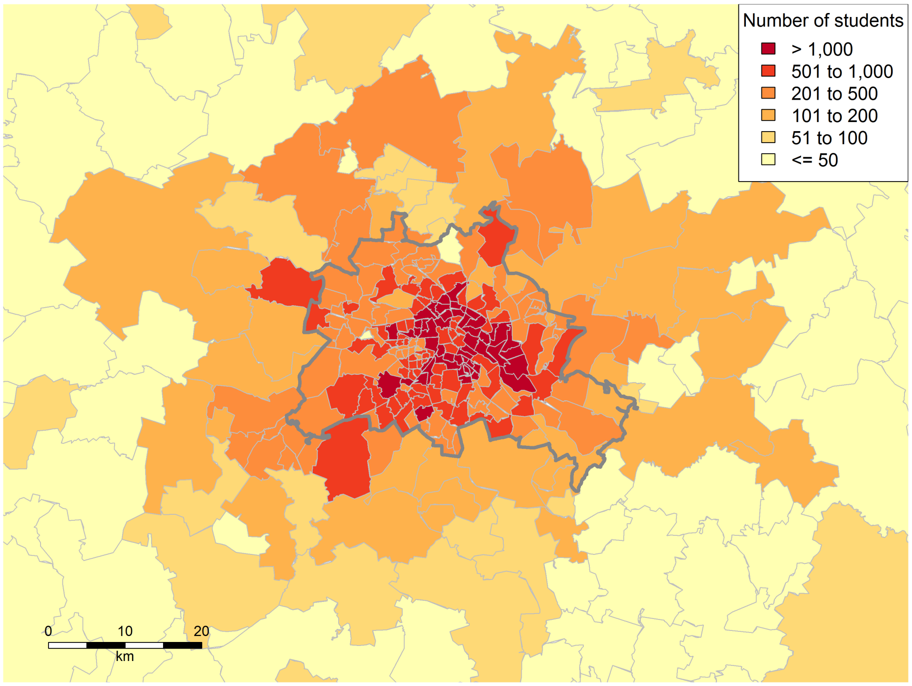

Figure 5: Distribution of Berlin students on ZIP-code area level in 2015.

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

When it comes to see the individual development of the LORs with the highest student counts one notices that they are located in special districts of Berlin (Wedding, Neukölln, Moabit, Prenzlauer Berg, Friedrichshain and Kreuzberg). Table 5 lists the ten most popular LORs areas among students (2015) and their development over time. They exhibit remarkable changes in their student population over time, thus restating the necessity of studies aiming to monitor the changes of the student population at a low level of regional aggregation. While the situation is quite stable in the districts Neukölln, Friedrichshain

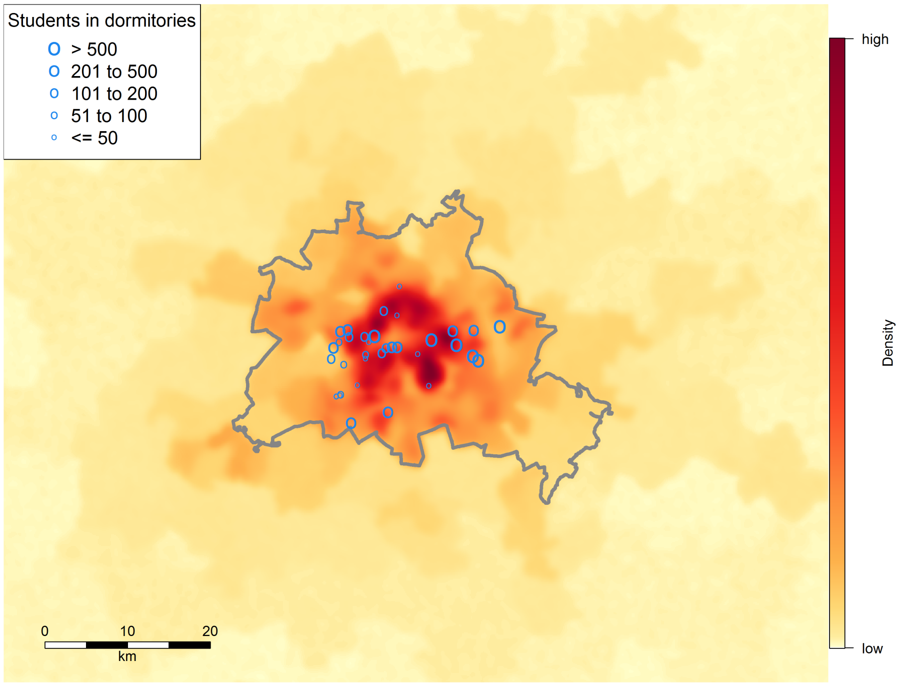

Figure 6: Kernel density estimates of Berlin students in 2015. Location of dormatories with its number of students is added.

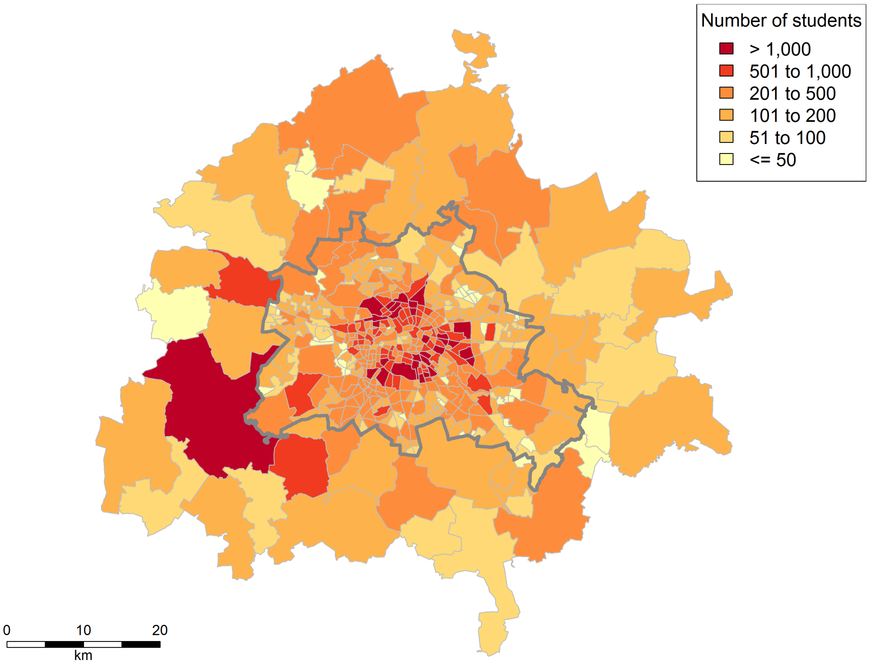

Figure 7: Distribution of Berlin students on administrative planning area level in 2015.

and Pankow there is a remarkable change towards higher student counts in the district of Wedding in the north of the central belt.

Table 5: The ten most popular urban planning areas LOR in 2015 with students counts for 2005, 2010 and 2015.

urban planning area (LOR)	central locality ('Ortsteil')	2015 count	2010 count	2005 count
Reuter Kiez	Neukölln	1938	2057	1943
Samariterviertel	Friedrichshain	1723	1774	1820
Rixdorf	Neukölln	1711	1425	770
Westhafen	Wedding	1595	1142	773
Rehberge	Wedding	1553	1082	726
Soldiner Straße	Wedding	1512	1006	680
Humboldthain Nordwest	Wedding	1467	1042	691
Reinickendorfer Straße	Wedding	1362	898	551
Pankow Süd	Pankow	1329	1297	1358
Emdener Straße	Moabit	1288	994	764

3.4.2 The temporal development 2005-2015

As a by product of the proposed routine one obtains the KDE maps for each of the three reference years 2005, 2010 and 2015 . These maps are displayed in Figure 8. The structure of the students settlement remains quite stable from this representation. However, if the display the highest densities regions ('HDR') we will notice remarkable regional changes. Note, however, that such a representation is restricted to the KDE approach.

Figure 9 compares the highest density regions ('HDR') 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. The parts of the eastern belt (southern Prenzlauer Berg and parts of Friedrichshain and Kreuzberg) did drop out of the 25% HDR in the last ten years. Interestingly it becomes apparent, that in general the concentration did decrease. The 25% highest density region enfolded only 24.64 km^2 in 2005. This area enlarged to 28.58 km^2 in 2010 and 33.27 km^2 in 2015. A similar effect is noticeable for the 50% HDR (2005: 76.88 km^2, 2010: 81.45 km^2, 2015: 92.40 km^2).

The observations described may be due to the general increase of student numbers by almost 50 % in Berlin. 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. This finding is consistent with the previous analysis of the Senate Department for Urban Development and Environment, showing a shortage and a price increase in the planning areas, which have lost their importance for student residency. 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.

Analysing 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 10.

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 were recorded for Prenzlauer Berg (Northeast) and the northern part of Mitte (Center), which can be attributed to the

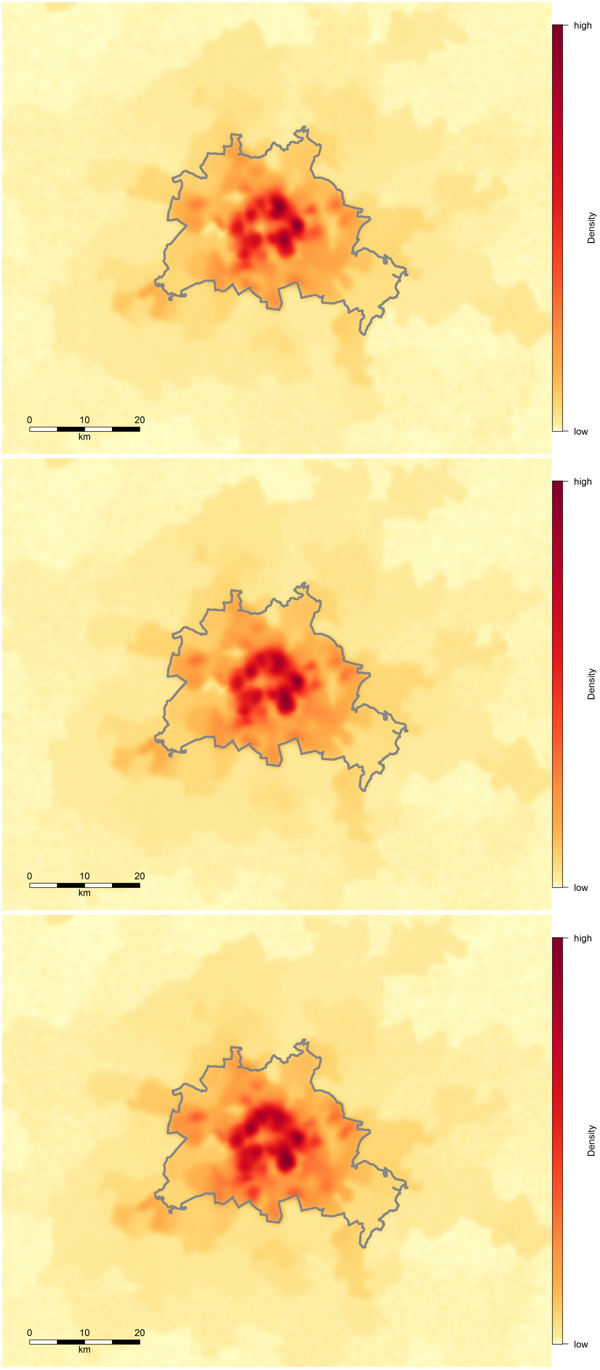

Figure 8: Kernel density estimates of students 2005 (upper), 2010 (middle) and 2015 (lower).

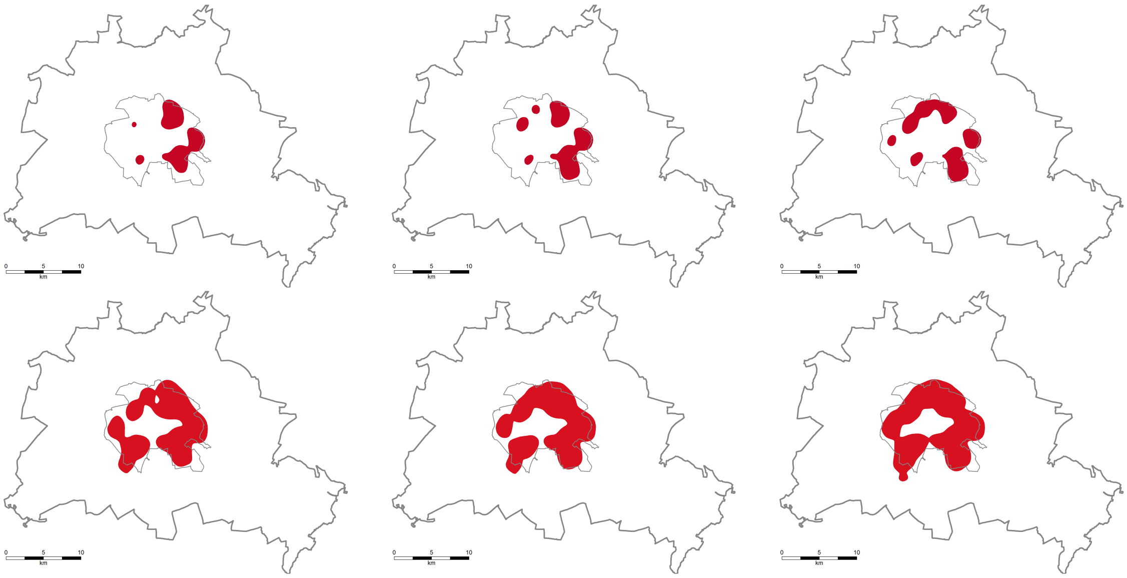

Figure 9: Regions with highest student density (left: 2005, middle: 2010, right: 2015). Upper panels: 25% of students. Lower panels: 50% of students.

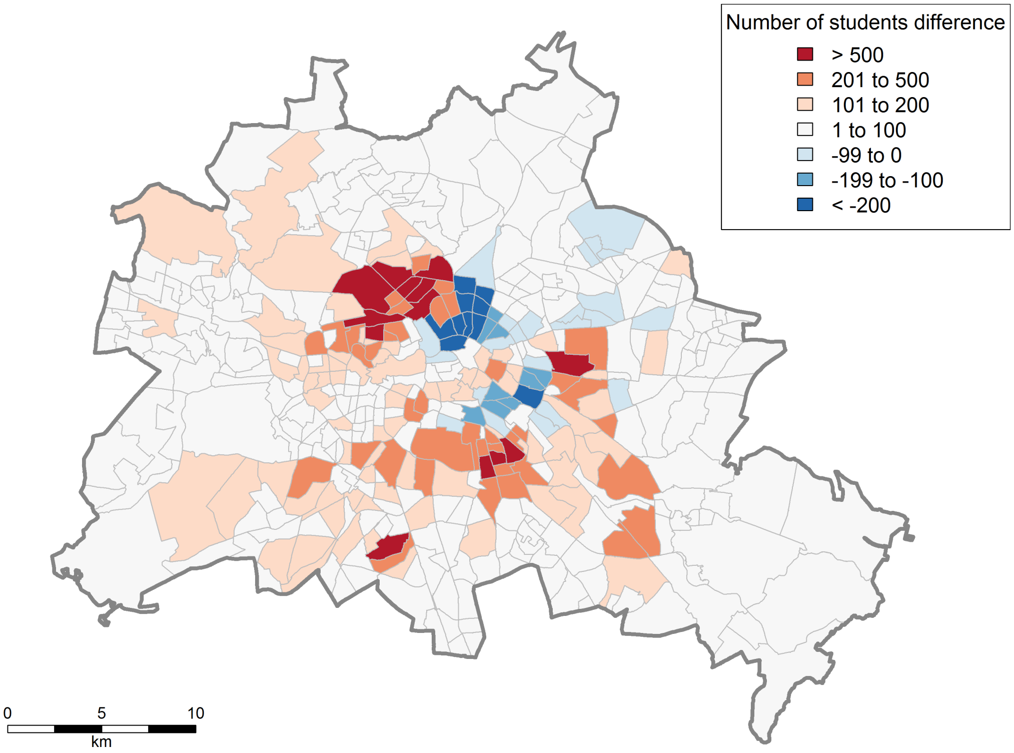

Figure 10: Differences in student numbers 2015 compared to 2005 on administrative planning area level.

gentrification of these quarters. In addition, the eastern parts of Friedrichshain (East) and Kreuzberg (Southeast) have lost students in the reference period.

4 Conclusion

Kernel density estimates are a useful tool in the business to transform case number between area systems which are not hierarchical. Even more, they are easier to use than ad-hoc solutions which base on unrealistic assumptions about uniform distributions of the characteristic of interest and are often carried out manually. With the free R-Package *Kernelheaping* the user can do this task quite automatically.

However, density estimates, which are used here as a transmission tool, have their own merits as they offer the display of highest density regions which can be used to identify local concentrations in the region of interest.

The simulation of quasi-exact geo-referenced data is a necessary tool for the estimation of the kernel densities. It turned out to be a useful tool also for the computation of case numbers in areas.

It should be noted that our algorithm is extremely useful for the construction of maps that are based on so-called ”open data”, see `https://en.wikipedia.org/wiki/Open_data`. Because of confidentiality reasons and their easy access they are often displayed as local aggregates. For example, in Berlin the open data are presented at the level of LORs or at a grid level, see `https://daten.berlin.de/datensaetze`.

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. A calibration is therefore 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 3). Beforehand, we also applied a correction for the HU and the FU in 2015 as their student numbers refer to the summer term instead of the winter term where student numbers are typically lower. Thus, we multiplied the numbers of these two universities by the ratio of winter term to summer term 2015 (e.g. FU: 36,674/33,173=1.11). 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 in the inner city borders but some are living near the college as well as the kernel density estimate for 2015 exhibits (cf. Figure 11).

For the year 2015 we think that the effect of missing colleges is negliglable as we have information on the most important ones and the remaining ones are rather small and quite similar distributed. If we would leave out the colleges only available in 2015 we get quite different area aggregates for ZIP-Codes near the missing colleges, e.g. ZIP-code 10318 with only 145 instead of 796 students. Figure 12 shows exemplary 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

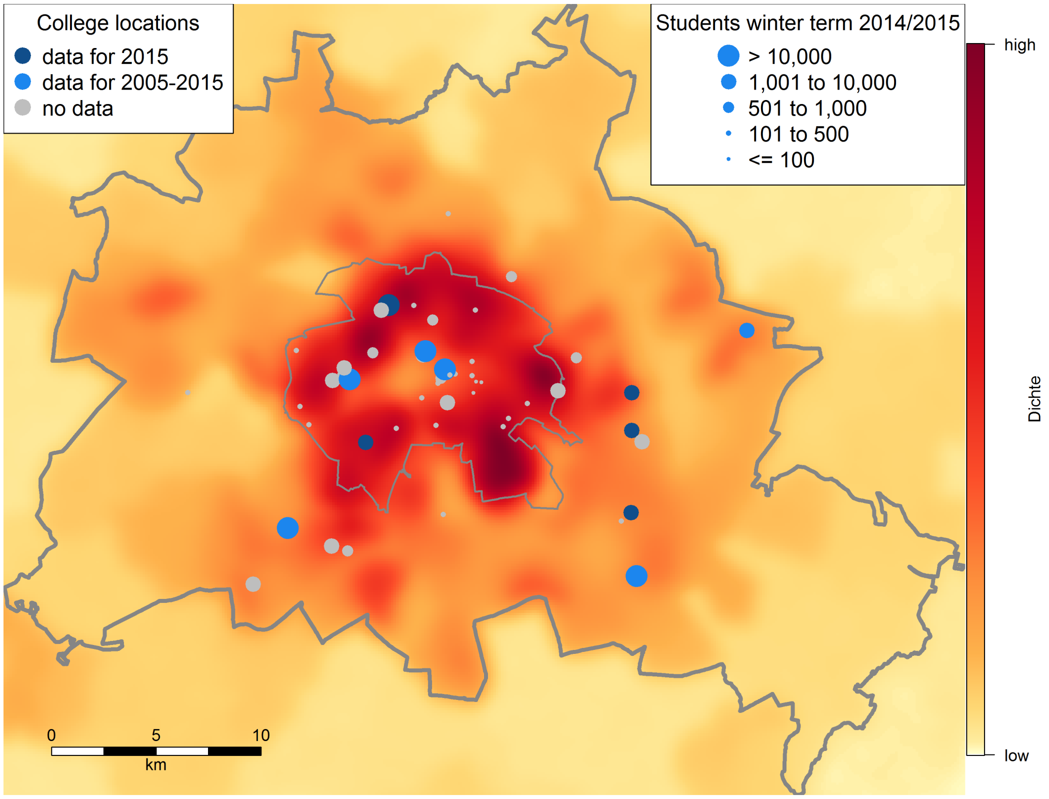

Figure 11: Locations of colleges and universities of Berlin with number of students including the kernel density estimate of the student distribution in 2015. The border of the 'inner city' is added to the map.

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.

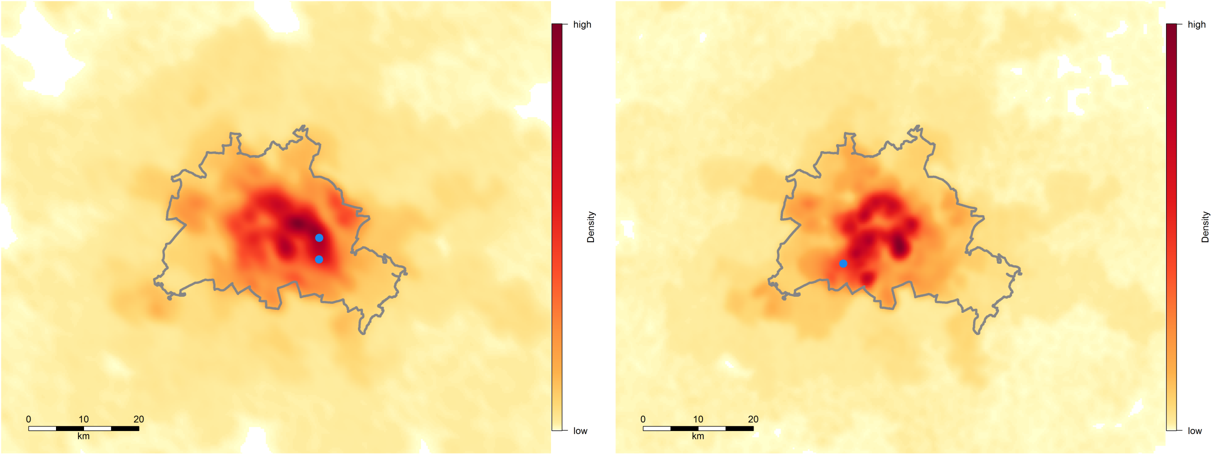

Figure 12: Kernel density estimates of HTW (left) and FU (right) student distributions with college site locations.

...

References

Celeux, G., Chauveau, D., and Diebolt, J. (1996), "Stochastic Versions of the EM algorithm: an Experimental Study in the Mixture Case," Journal of Statistical Computation and Simulation, 55(4), 287–314.

Groß, M. (2016), Kernelheaping: Kernel Density Estimation for Heaped Data. R package version 1.6.

Groß, M., Rendtel, U., Schmid, T., Schmon, S., and Tzavidis, N. (2016), "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 (Statistics in Society), p. forthcoming.

McCulloch, C. E., and Neuhaus, J. M. (2001), Generalized linear mixed models Wiley Online Library.

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

Diskussionsbeiträge - Fachbereich Wirtschaftswissenschaft - Freie Universität Berlin
Discussion Paper - School of Business and Economics - Freie Universität Berlin

2018 erschienen:

2018/1 BESTER, Helmut und Ouyang YAOFU
Optimal Procurement of a Credence Good under Limited Liability
Economics