REGIONALANALYSE AUF BASIS SIMULIERTER GEOKOORDINATEN

Gütebeurteilung des Verfahrens am Beispiel der Wahlberechtigten in Berlin

Kerstin Erfurth

Schlüsselwörter: Kerndichteschätzung – Simulierte Geokoordinaten – Kernelheaping – Choroplethenkarten – regionale Analyse

ZUSAMMENFASSUNG

Für Daten mit geografischem Bezug eignen sich Kartendarstellungen zur Visualisierung, um einen einfachen Zugang zu komplexen Informationen zu erhalten. Insbesondere die Verteilung verschiedener Bevölkerungsgruppen und die Identifikation von Hotspots stellen ein für Planungszwecke bedeutendes Interesse dar. Diese Arbeit beschäftigt sich mit der Bewertung des neuen Kernelheaping-Verfahrens gegenüber anderen in der Praxis gängigen Verfahren zur kartografischen Dichteschätzung von Daten. Dazu wurde ein praxisnahes Szenario mit den Daten der Wahlberechtigten in Berlin geschaffen, in welchem unter kontrollierten Bedingungen Vergleiche durchgeführt werden können. Es konnte gezeigt werden, dass das Kernelheaping-Verfahren in der Lage ist, qualitativ bessere Ergebnisse zu erzielen als die bisher verwendeten Standardverfahren.

Keywords: *kernel density estimation – simulated geo-coordinates – kernelheaping – choropleth maps – regional analysis*

ABSTRACT

For data with a spatial reference, map representations are suitable for visualisation in order to gain easy access to complex information. In particular the distribution of different population groups and the identification of hotspots represent a major interest for planning purposes. This work deals with the evaluation of the new kernelheaping method compared to other methods used in practice for the cartographic density estimation of data. For this purpose, a practical scenario was created with Berlin's voter data, in which comparisons can be carried out under controlled conditions. It was shown that the kernelheaping procedure is able to achieve results of better quality than the standard methods used to date.

Kerstin Erfurth

hat Statistik (M. Sc.) an der Humboldt-Universität zu Berlin studiert und zusätzlich das Zertifikat European Master in Official Statistics (EMOS) erlangt. Im Rahmen des EMOS-Programms absolvierte sie einen Forschungsaufenthalt im Amt für Statistik Berlin-Brandenburg. Aus dieser Kooperation ist ihre Masterarbeit zum Thema „Gütebeurteilung und Einsatz simulierter Geokoordinaten bei der regionalen Analyse zur Bundestagswahl 2017" entstanden, für die sie mit dem Gerhard-Fürst-Preis 2019 in der Kategorie „Master-/Bachelorarbeiten" ausgezeichnet wurde. Seit November 2018 arbeitet sie im Amt für Statistik Berlin-Brandenburg.

1

Einführung

Die Erhebung und Interpretation von Daten ist zu einem zentralen Thema in der modernen Informationsgesellschaft geworden. Ein einfacher Zugang zu komplexen Informationen wird durch geeignete Visualisierungen ermöglicht, wobei die gewählte Darstellungsmethode einen wesentlichen Einfluss auf die Interpretation der Daten haben kann. Für Daten mit geografischem Bezug eignen sich insbesondere Kartendarstellungen, welche unter anderem mit farbigen Symbolen, Grenzlinien oder Flächen angereichert werden, um deren Raumbeziehungen und ihre relativen Verhältnisse leicht verständlich zu machen. Dabei spielen auch Diskretisierung[1], Kategorisierung und die verwendeten Farbabstufungen eine große Rolle, da der visuelle Eindruck durch deren Wahl stark beeinflusst werden kann. Im Umgang mit aggregierten Daten wird der Ansatz der Vorverarbeitung zu einer Schlüsseltechnik für gute Ergebnisse. Um umfassende Informationen über alle interessierenden Geokoordinaten zu erhalten, wird ein neuer nicht parametrischer Ansatz zur Dichteschätzung namens Kernelheaping evaluiert.

Der neue Ansatz wird statistisch mit einer zugrunde liegenden bekannten realen Dichte von Wahlberechtigten in Berlin auf verschiedenen Aggregationsebenen bewertet und quantitativ den Standardverfahren vergleichend gegenübergestellt. Dafür konnte auf anonymisierte Adressdaten des Amtes für Statistik Berlin-Brandenburg zugegriffen werden.

Es wird gezeigt, dass das Kernelheaping von den untersuchten Verfahren die beste Möglichkeit bietet, lokale Aggregate unabhängig von Verwaltungsgrenzen, zum Beispiel Wahlbezirken, zu behandeln. Es ermöglicht geografische Kartendarstellungen mit exakten Geokoordinaten, welche als „Wähler je Pixel“ interpretiert werden können, auch wenn die exakten Geokoordinaten für Wahldaten ursprünglich nicht zur Verfügung stehen. Dies dient neben der Visualisierung von Verteilungen auch der Identifikation von Hotspots interessierender Personengruppen.

1 Diskretisierung beschreibt die Zerlegung stetiger, räumlicher Flächen in kleine Abschnitte oder Punkte.

2

Datengrundlage „Adressdichte“

Als Analysegrundlage werden anonymisierte Einzeldaten zur Anzahl der Wahlberechtigten in Berlin auf Adressebene genutzt. Sie stellt eine Grundgesamtheit dar, da alle Personen erhoben wurden, die im Dezember 2016 ihren Hauptwohnsitz in Berlin gemeldet hatten. Der kritische Aspekt ist demzufolge nicht der Informationsverlust durch die Ziehung einer Stichprobe, von welcher auf die Grundgesamtheit geschlossen werden soll. Der Fokus liegt auf dem Informationsverlust durch die Aggregation der Daten. Es werden prinzipiell alle Personen erfasst, ihre räumlichen Geokoordinaten werden jedoch gerundet beziehungsweise aggregiert. Daher ist die Anwendung insbesondere für amtliche Daten interessant, beispielsweise für die in dieser Arbeit untersuchten Wahlberechtigtenzahlen.

Voraussetzung für nicht parametrische Ansätze für Dichteschätzungen ist die „Glattheit“ der zu schätzenden Dichte (Fahrmeir und andere, 1996). Aus diesem Grund wird über die Adressdaten eine minimale Kerndichte als Glättungsprozedur gelegt. Auf diese Weise wird eine geglättete Version der Originaldaten erzeugt, welche nicht nur die Voraussetzungen für Dichteschätzungen erfüllt, sondern auch bessere Visualisierungsmöglichkeiten liefert. Wird der Volumeninhalt unter der bivariaten Dichte-Kurve auf Eins normiert, entsteht die „Adressdichte“ der Wahlberechtigten. Diese Adressdichte ist die Ausgangsbasis der Untersuchungen und stellt eine sehr realitätsnahe Datengrundlage einer „wahren Dichte“ dar.

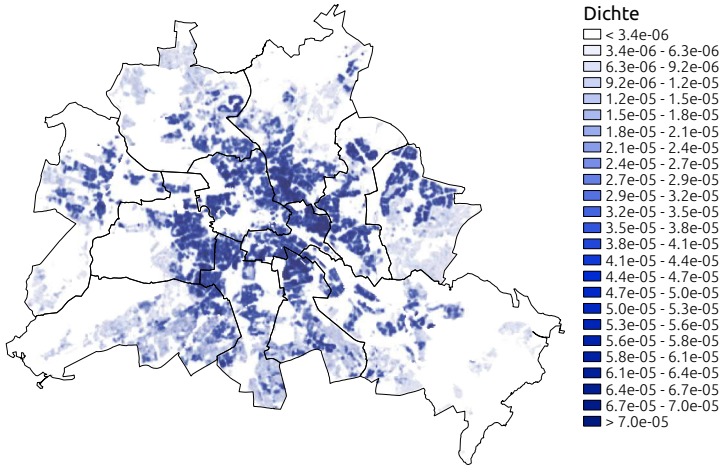
Um zu einer numerisch verarbeitbaren Datenbasis zu gelangen, muss darüber hinaus eine Diskretisierung in Pixel durchgeführt werden. Auch zu Darstellungszwecken ist dieser Schritt unvermeidbar. Dabei wird jedem Pixel ein Dichtewert zugeordnet, sodass unter Berücksichtigung der Pixelgröße der Flächeninhalt unter allen Pixeln ebenfalls Eins ergibt. Die diskretisierte Version der Adressdichte beschreibt demzufolge wieder eine Dichte.

Diese Vorverarbeitungsschritte liefern die finale Adressdichte auf Pixelbasis. Dabei wurde in den Untersuchungen die Größe der Pixel so gewählt, dass sie etwa einem Hektar Landfläche entspricht. Bei einer Umrechnung

Grafik 1
Umrechnung auf die finale Adressdichte auf Pixelbasis

Verteilung der Berliner Adressen über das Land mit exakten Geokoordinaten

Kartografische Darstellung der Adressdichte mit einheitlicher Skala



Datenquelle: Anonymisierte Adressdaten des Amtes für Statistik Berlin-Brandenburg, Stichtag 31.12.2016

2020 - 01 - 0162

des Dichtewertes auf Wahlberechtigte entsteht eine leicht zu interpretierende Größe „Wahlberechtigte je Hektar“. Diese Adressdichte wird in allen Berechnungen als wahre Dichte angenommen und genutzt, um Fehlerterme und Gütekriterien zu berechnen. Alle untersuchten Dichteschätzungen werden auf das gleiche Raster zurückgeführt, um eine Vergleichbarkeit herzustellen. ↘ Grafik 1

3

Aggregationslevel

Um Berechnungen auf Basis unterschiedlicher Aggregationsstufen kontrolliert durchführen zu können, werden die ursprünglichen Wahlberechtigtenzahlen zunächst auf acht vorab gewählte regionale Bezugssysteme kumuliert. Auf diese aggregierten Daten werden die verschiedenen Verfahren angewendet und anschließend verglichen. Je mehr Polygonzüge ein Aggregationslevel besitzt, desto detaillierter ist es und desto geringer ist der Informationsverlust im Vergleich zu den Originaldaten. In der Untersuchung wurden die zwölf Berliner Bezirke (BEZ), 60 Prognoseräume (PRG), 96 Ortsteile (ORT), 138 Bezirksregionen (BZR), 192 Postleitzahlen (PLZ), 447 Planungsräume (PLR), 660 Briefwahlbezirke (BWB) und 1 779 Urnenwahlbezirke (UWB) verwendet. Diese Strukturen sind zum Teil natürlich gewachsen und liegen nicht in einem regelmäßigen Raster oder Gitter. Teilweise sind die Flächen unterschiedlich groß mit stark variierenden Einwohnerzahlen. Derartige Inhomogenitäten treten sehr häufig bei administrativen Grenzen auf. Mit der Betrachtung verschiedener Aggregationsstufen kann abhängig von den gewählten Verfahren festgestellt werden, ab welchem Grad des Informationsverlusts in den Daten eine Anwendung überhaupt noch sinnvoll sein kann.

4

Verfahren

Die Verfahren, welche zentral verglichen werden, sind Choroplethenkarten, klassische (naive) Kerndichteschätzung und das Kernelheaping-Verfahren für simulierte Geokoordinaten.

Die im Fokus stehende Methode ist dabei der Kernelheaping-Algorithmus. Er erzeugt iterativ eine Kerndichte und berücksichtigt damit die Tatsache, dass regional aggregierte Daten vorliegen. Dieser Algorithmus ist eine Anwendung eines Stochastic-Expectation-Maximization(SEM)-Algorithmus, bei dem der stochastische Teil als geschichtete Stichprobe von Geokoordinaten

aus der Dichte des vorherigen Schrittes realisiert wird. Eine Schicht entspricht dabei einem Polygon. Nach einigen Iterationen konvergiert der Algorithmus zu einer Dichte mit Schätzwerten für alle Geokoordinaten. Für eine erfolgreiche Anwendung ist eine feste, aber frei wählbare Anzahl von Iterationen erforderlich. In der Analyse wird dieser Parameter systematisch modifiziert, um Unterschiede in den Ergebnissen aufzuzeigen und optimale Parametereinstellungen zu finden.

Zusätzlich werden klassische Choroplethenkarten auf die Datensätze angewendet. Diese sind Standard für die Visualisierung von aggregierten Datensätzen in der amtlichen Statistik. Für Choroplethenkarten sind keine genauen Geokoordinaten erforderlich. Ihr größter Nachteil ist die homogene Farbe innerhalb der regionalen Einheiten. In der Regel sind die Farbkategorien auf etwa fünf Ebenen begrenzt, was einen erheblichen Informationsverlust bedeutet. Darüber hinaus haben regionale Einheiten im Allgemeinen nicht die gleiche Größe. Daher kann die Interpretation von Choroplethenkarten über die Flächengrößen, welche am attraktivsten ist, zu irreführenden Schlussfolgerungen führen. In der Analyse wird gezeigt, dass einfache Flächennormalisierungen bereits helfen, realistischere Kartendarstellungen zu erstellen.

Für Vergleichszwecke wird zudem ein naives Kerndichte-Schätzverfahren (kernel density estimation – KDE) verwendet. Es handelt sich um eine gängige Glättungstechnik, mit der eine Dichte für aggregierte Daten geschätzt werden kann. Ein wesentlicher Nachteil dieses Verfahrens ist die komplexe Aufgabe, geeignete Glättungsparameter (Bandbreiten) zu finden.

Für alle Methoden wurden systematisch Parameter angepasst, um optimale Einstellungen aufzuzeigen und die Techniken hinsichtlich ihrer Robustheit zu bewerten.

5

Diskretisierung

Wie oben erwähnt, wurden alle Dichten in gleicher Art und Weise diskretisiert, um eine einheitliche Berechnungsgrundlage zu schaffen. Dabei wurde für alle Methoden eine identische Rastergröße (Gridsize) festgelegt, welche eine interpretierbare Pixelgröße liefert. Aus diesem Grund ist die Gridsize selbst zunächst eine nicht intuitiv krumme Zahl. Berlin besitzt eine Ost-West-Ausdehnung von etwa 45 784 Metern. Sollen Pixel mit einer Fläche von einem Hektar entstehen, so ist eine Gridsize von 458 Pixel in horizontaler Richtung sinnvoll. Die Nord-Süd-Spanne Berlins beträgt etwa 37 739 Meter. Da die entstehenden Pixel quadratisch sein sollen, ergibt sich eine Gridsize in vertikaler Ausrichtung von 378 Pixel. Alle finalen Karten besitzen demzufolge eine Auflösung von 458 x 378 Pixel.

6

Bewertungskriterien

Um schließlich Vergleiche auf der Grundlage quantitativer Kriterien mithilfe der realen Daten zu ermöglichen, wurde der mittlere quadratische Fehler (mean squared error – MSE) berechnet. Dieser basiert auf dem Bias und der Varianz. Alle drei Werte können in einem ersten Schritt pixelweise erhoben werden, weil die Ergebnisse in derselben Art und Weise diskretisiert vorliegen und die wahre Dichte (Adressdichte) bekannt ist. Da sich eine Auswertung auf Pixelbasis gut für eine grafische Darstellung eignet, aber einen Vergleich zwischen den Verfahren erschwert, werden in einem zweiten Schritt über alle Pixel gemittelte Werte bestimmt. Für die Choroplethenkarten und die naive Kerndichteschätzung fehlt durch die deterministische Berechnung der zufällige Anteil, weswegen es für diese Schätzungen keine Verfahrensvarianz gibt. Dennoch werden die entsprechenden Bewertungskriterien für die Choroplethenkarten und die naive Kerndichteschätzung analog ermittelt. Die Besonderheit dabei ist jedoch, dass die Verfahrensvarianz mit Null in die finale Kriterienberechnung eingeht.

Das Kernelheaping-Verfahren muss zur Bestimmung des Bias, der Varianz und des MSE mehrfach ausgeführt werden, da durch den stochastischen Anteil das Ergebnis jeder Ausführung leicht variiert. Mit der für die Analysen implementierten Erweiterung des Verfahrens um unabhängige Berechnungsketten können diese direkt parallel innerhalb des Kernelheaping-Pakets berechnet werden. Auf diese Weise entstehen mehrere voneinander unabhängige Markov-Ketten[2] je Pixel. Es wird eine

2 Markov-Ketten sind besondere stochastische Prozesse, bei denen der zukünftige Zustand eines Prozesses nur durch den aktuellen Zustand bedingt wird. Er wird nicht durch vergangene Zustände beeinflusst.

initiale „Burnin“-Phase definiert, deren Iterationsergebnisse im späteren Verlauf verworfen werden. Danach folgen „Sample“-Iterationen, deren Dichteschätzungen in das Endergebnis einfließen. Um für eine Kette zu einem Ergebnis für ein Pixel zu gelangen, wird über alle Iterationen der Mittelwert gebildet. So entsteht für jede Markov-Kette eine separate Verteilung, welche einen Mittelwert und eine Varianz aufweist.

Dieses Prinzip lässt sich entsprechend auch über alle Ketten anwenden, um zu einem Gesamtmittelwert zu gelangen, der dem finalen Schätzergebnis je Pixel entspricht. Darüber hinaus ergibt sich eine Verfahrensvarianz, die durch die mehrfache Durchführung des Verfahrens entsteht und empirisch ermittelt werden kann.

7

Ergebnisse

In den Analysen wurden alle Verfahren für alle Aggregationsstufen berechnet und kartografisch dargestellt. Die folgenden Grafiken zeigen ausgewählte Karten, in denen immer Bezug auf die Dichtewerte der einzelnen Pixel genommen wird. Auf eine Umrechnung der „Dichtewerte je Pixel“ auf „Wahlberechtigte je Hektar“ wird an dieser Stelle verzichtet. Diese Skalierung dient einer intuitiven Interpretation, ist jedoch für Vergleichsuntersuchungen nicht notwendig.

7.1 Choroplethenkarten

Die Berechnung von Choroplethenkarten basiert zunächst auf der einfachen Idee, dass Wahlberechtigte innerhalb eines Bezirks je nach Aggregationslevel aufsummiert werden und sich die regionale Zuordnung ausschließlich auf diesen Bezirk bezieht. So ergibt sich je Bezirk eine Gesamtzahl an Wahlberechtigten. Bei dieser Darstellung geht der exakte räumliche Bezug eines jeden Wahlberechtigten verloren. Es kann keiner Person eine genaue Geokoordinate zugeordnet werden, stattdessen nur noch ein Bezirk.

Um die Choroplethenkarte als nicht parametrische Dichteschätzmethode in die Vergleichsuntersuchungen aufnehmen zu können, muss nach der Aggregation der Wahlberechtigten eine Normierung stattfinden. So ist die Choroplethenkarte gegenüber der Originaldarstellung zwar anders skaliert, das Verhältnis zwischen den Daten bleibt jedoch erhalten. Damit hinterlässt eine Karte den gleichen optischen Eindruck.

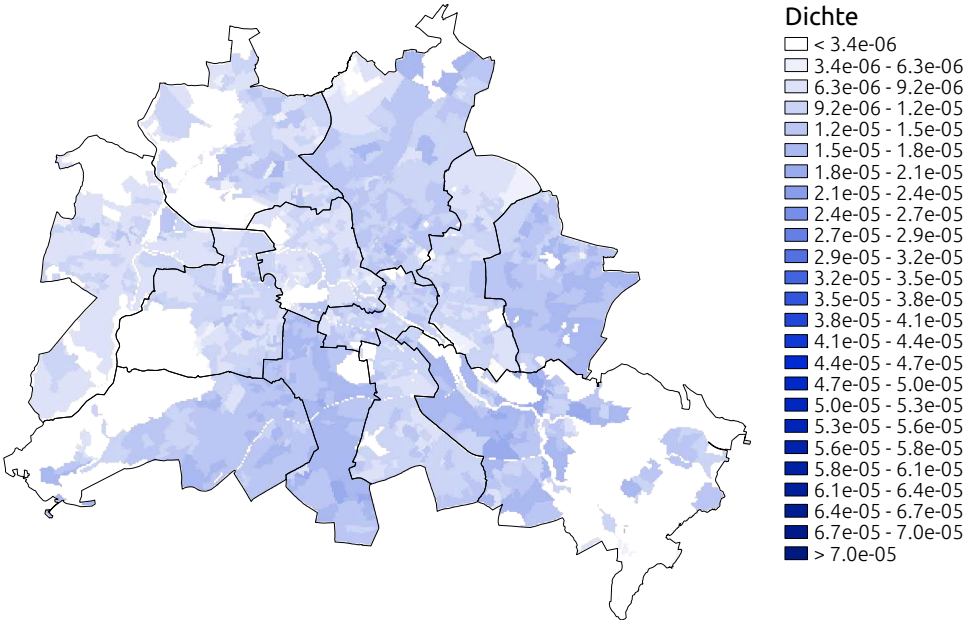
Es gibt einen weiteren Ansatz, die Choroplethenkarte zu normieren, um zu einer Dichte zu gelangen. Bei einer einfachen Normierung wird mithilfe der Gesamtzahl der Wahlberechtigten aller Bezirke skaliert. Um die inhomogenen Flächengrößen der verschiedenen Bezirke zu berücksichtigen, wird die Anzahl der Wahlberechtigten je Bezirk zunächst durch die jeweilige Bezirksgröße geteilt und anschließend die Gesamtheit auf Eins normiert. Auf diese Weise entsteht eine Dichte, welche einen Wahlberechtigtenanteil je Flächeneinheit (hier Pixel) zulässt. So werden die verschiedenen Bezirksgrößen berücksichtigt und es entsteht ein adäquater Vergleich zu Kerndichten. Beide Verfahren wurden in die Analysen einbezogen.

Die Schätzergebnisse der Choroplethenkarten zeigen die folgenden Grafiken 2 und 3. Bei ↘ Grafik 2 handelt es sich (abgesehen von der Normierung) um eine klassische Darstellung der Verteilung von Wahlberechtigten auf Urnenwahlbezirke. Die simple Choroplethenkarte hat einen auffallend geringen Kontrast. Die Flächen werden insbesondere für die Wahlbezirke so geschnitten, dass ähnliche Wählerzahlen entstehen. Auf diese Weise wird der Bearbeitungsaufwand je Bezirk vergleichbar gehalten. Bei der Visualisierung der Wahlberechtigten entsteht in der Farbgebung daher ein homogenes Gesamtbild. Die Flächengröße bleibt dabei unberücksichtigt. Dies ist das Hauptproblem der simplen Choroplethenkarte.

Aus diesem Grund ist in ↘ Grafik 3 mit der flächennormierten Choroplethenkarte eine Anpassung an diese Problematik dargestellt. Dieses Vorgehen liefert eine differenziertere Darstellung der Wahlberechtigtendichte. Dennoch gibt es keine Abstufungen innerhalb eines Bezirks, was insbesondere an den Bezirksgrenzen sichtbar wird. Im Ortsteil Moabit in Berlin-Mitte stößt beispielsweise eine sehr helle Fläche auf eine sehr dunkle Fläche. Dieser Effekt spiegelt die Realität nur unzureichend wider. Obwohl die Flächengröße einbezogen wird, sind die Informationen nur flächenbezogen abgebildet. Dem gegenüber stehen die punktbezogenen Schätzungen, die eine deutlich genauere Auflösung der Ergebnisse erlauben.

Grafik 2

Kartografische Darstellung der Ergebnisse für die simple Choroplethenkarte

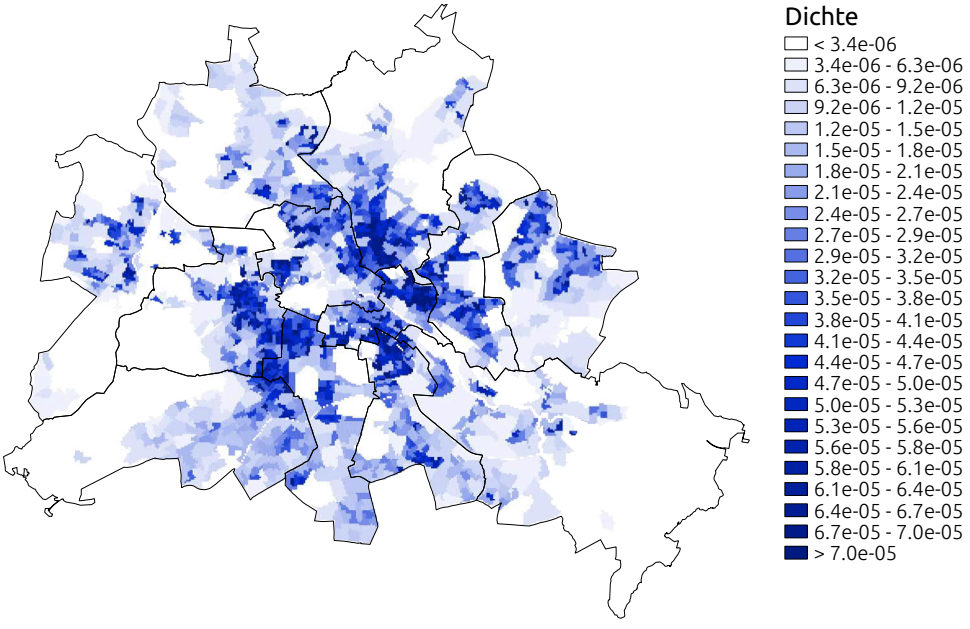


Dargestellt ist die Verteilung von Wahlberechtigten auf Urnenwahlbezirke in Berlin zum Stichtag 31.12.2016.

2020 - 01 - 0163

Grafik 3

Kartografische Darstellung der Ergebnisse für die flächennormierte Choroplethenkarte



Dargestellt ist die Verteilung von Wahlberechtigten auf Urnenwahlbezirke in Berlin zum Stichtag 31.12.2016.

2020 - 01 - 0164

7.2 Naive Kerndichteschätzung

Für die Berechnungen der naiven Kerndichteschätzungen wurde das in R implementierte Paket ks von Tarn Duong (Duong, 2017) verwendet. Dieses entspricht in seiner Funktionsweise den gängigen, in Geoinformationssystemen verfügbaren Plugins zur Kerndichteschätzung. Bei der naiven Kerndichteschätzung handelt es sich im Gegensatz zur Choroplethenkarte um ein punktbezogenes Verfahren zur Dichteschätzung.

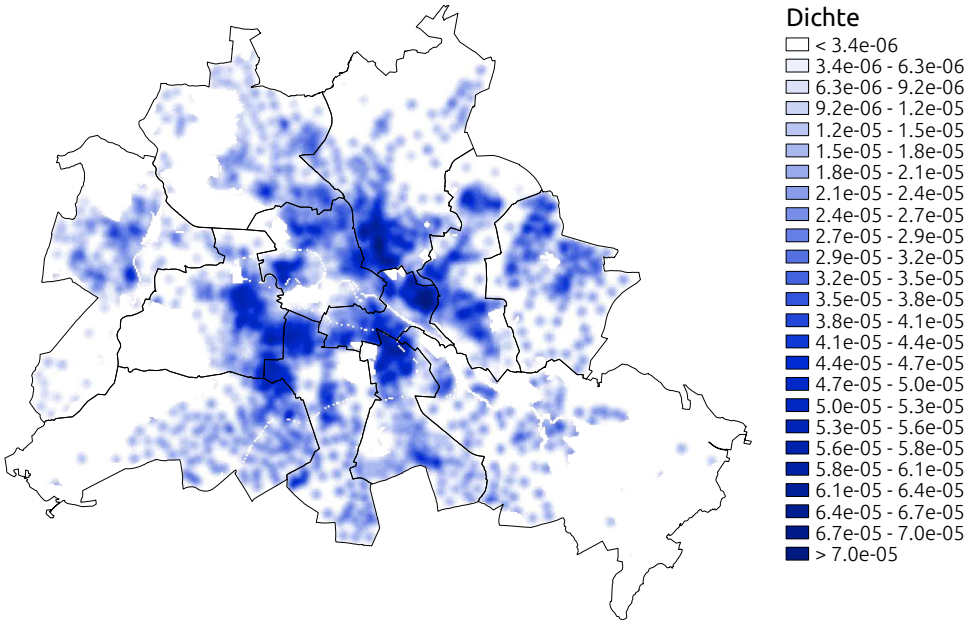
Durch den Aggregationsprozess entsteht ein Informationsverlust von Geokoordinaten. Da diese jedoch für eine Schätzung benötigt werden, muss für das Verfahren initial jeweils ein beliebiger Punkt innerhalb eines Bezirks gewählt werden. Für die in den Analysen durchgeführten Berechnungen werden die geografischen Mittelpunkte der Polygone genutzt. Darauf wird eine einfache Glättung angewendet. Darüber hinaus bleibt eine wesentliche Entscheidungsmöglichkeit innerhalb des Verfahrens die Wahl der Bandbreite.

Für den finalen Verfahrensvergleich wurden zwei verschiedene Varianten für die Wahl der Bandbreite einbezogen. Zum einen wurde die Bandbreite optimal anhand des Datensatzes selbst gewählt. Dabei gilt die Bandbreite als optimal, welche im Vergleich zur wahren Adressdichte den minimalen MSE liefert. Dies ist nur möglich, wenn eine wahre Dichtekarte bekannt ist. Zum anderen – da die wahre Dichte in der Praxis im Allgemeinen nicht zur Verfügung steht – wurde der Plugin-Selektor nach Wand und Jones (Wand/Jones, 1994) des R-Pakets ks genutzt. Dieser soll eine optimale Bandbreite liefern. Dabei entstehen auf jedem Aggregationslevel zwei verschiedene optimale Bandbreiten und entsprechend auch zwei verschiedene Dichteschätzungen, die sich in ihrer Erscheinung und Qualität unterscheiden.

↘ Grafik 4 zeigt die mittels der wahren Dichte berechnete naive Kerndichteschätzung mit optimaler Bandbreite. Dieses Ergebnis sieht optisch vielversprechend aus. Auffallend sind jedoch die vielen „Punktwolken“, welche durch die Wahl der geografischen Mittelpunkte entstehen. Die Bandbreite für die gesamte Karte wird einheitlich berechnet, daher ist sie in den größeren Randbezirken tendenziell zu klein gewählt, sodass Dichtepunkte nur im Zentrum der einzelnen Randbezirke auftreten.

Grafik 4

Kartografische Darstellung für die naive Kerndichteschätzung mit optimaler Bandbreite



Dargestellt ist die Verteilung von Wahlberechtigten auf Urnenwahlbezirke in Berlin zum Stichtag 31.12.2016.

2020 - 01 - 0165

Eventuell wird die flächennormierte Choroplethenkarte des vorherigen Abschnitts als optisch ansprechender empfunden, da diese scharfe Kanten besitzt, an denen sich das Auge festhalten kann. Die harten Kanten, wie sie beispielsweise in Moabit zu finden sind, geben jedoch kein wirklichkeitsnahes Szenario wieder. Im Vergleich zur wahren Dichte wirkt die naive Kerndichteschätzung daher bereits realistischer. Es ist jedoch zu beachten, dass für Grafik 4 die wahre Dichte zur Bandbreitenwahl genutzt wurde. Diese steht in der Praxis nicht zur Verfügung.

Das Schätzergebnis unter Verwendung eines Plugin-Verfahrens zur Bandbreitenwahl ist in ↘ Grafik 5 dargestellt. Hier fällt auf, dass viele Detailinformationen verloren gehen. Dies ist selbst auf Ebene der Urnenwahlbezirke ersichtlich, obwohl die Daten mit 1 779 Messwerten vergleichsweise genau erhoben wurden. Bei der Anwendung der naiven Kerndichteschätzung auf Stichproben für Rückschlüsse auf eine Grundgesamtheit wäre dieses Ergebnis plausibel, da die Grundstruktur der Verteilung prinzipiell gut wiedergegeben wird. In diesem Anwendungsfall jedoch entsteht durch die großflächigen Außenbezirke eine zu große Bandbreite, was zu dem starken Glättungseffekt führt.

7.3 Kernelheaping-Verfahren

Für die Berechnungen des Kernelheaping-Verfahrens wurde aus dem Paket kernelheaping von Marcus Groß (Groß, 2017) die Methode dshapebivr verwendet. Der Algorithmus durchläuft die folgenden Schritte:

Initialisierung

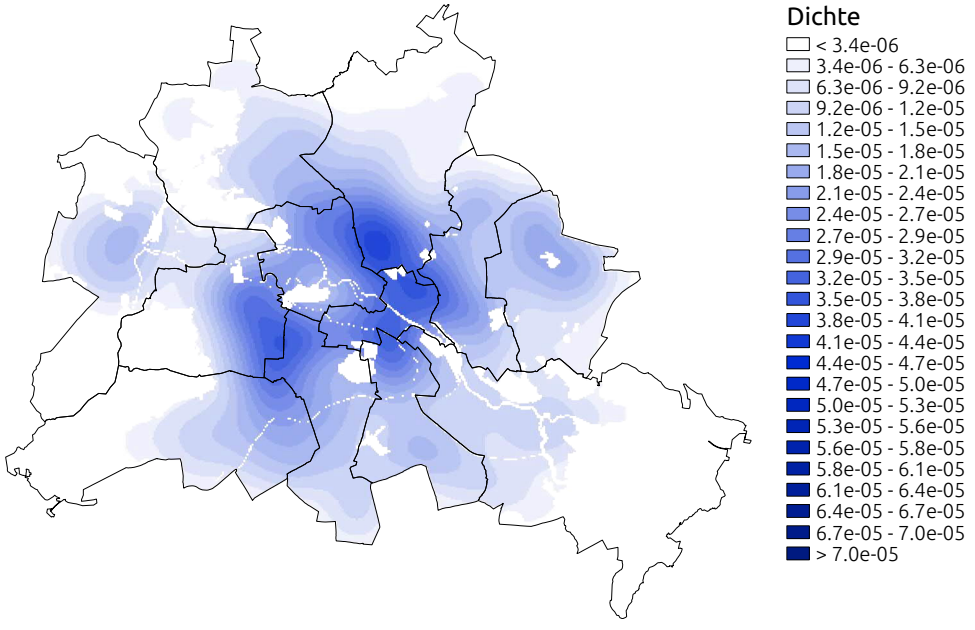
1. Berechnung der Anzahl der Pixel und ihrer zugehörigen Geokoordinaten basierend auf der Gridsize, um ein Raster zu erhalten, welches alle Bezirke der Aggregationsstufe abdeckt (Bounding-Box).
2. Berechnung einer naiven Kerndichteschätzung f_0 für alle Pixel mit den auf die jeweiligen Mittelpunkte aggregierten Daten mit einer initialen Bandbreite, basierend auf der Größe der Bounding-Box und der Anzahl der Bezirke.

Iteration

Führe Schritte 3. bis 6. für eine festgelegte Anzahl von Iterationen aus: $t = 1, \ldots, i$

3. Zufällige Ziehung von j_R Pixeln mit Zurücklegen für jeden Bezirk R basierend auf der Dichte f_t. Dabei ent-

Grafik 5
Kartografische Darstellung für die naive Kerndichteschätzung mit Plugin-Bandbreite



Dargestellt ist die Verteilung von Wahlberechtigten auf Urnenwahlbezirke in Berlin zum Stichtag 31.12.2016.

2020 - 01 - 0166

spricht die Anzahl der gezogenen Pixel dem zugehörigen Absolutwert der initial aggregierten Daten.

4. Berechnung der „optimalen“ Bandbreite nach Wand und Jones auf Basis der gezogenen Pixel
5. Berechnung einer naiven Kerndichteschätzung f_{t+1} auf Basis der gezogenen Pixel mit „optimaler“ Bandbreite
6. Nullsetzung der Dichtewerte aller Pixel außerhalb von Polygonen und ausmaskierten Pixel

Resultat

7. Berechnung eines Dichte-Mittelwertes für alle Pixel über die Dichten f_b, … , $f_{i\text{-}1}$

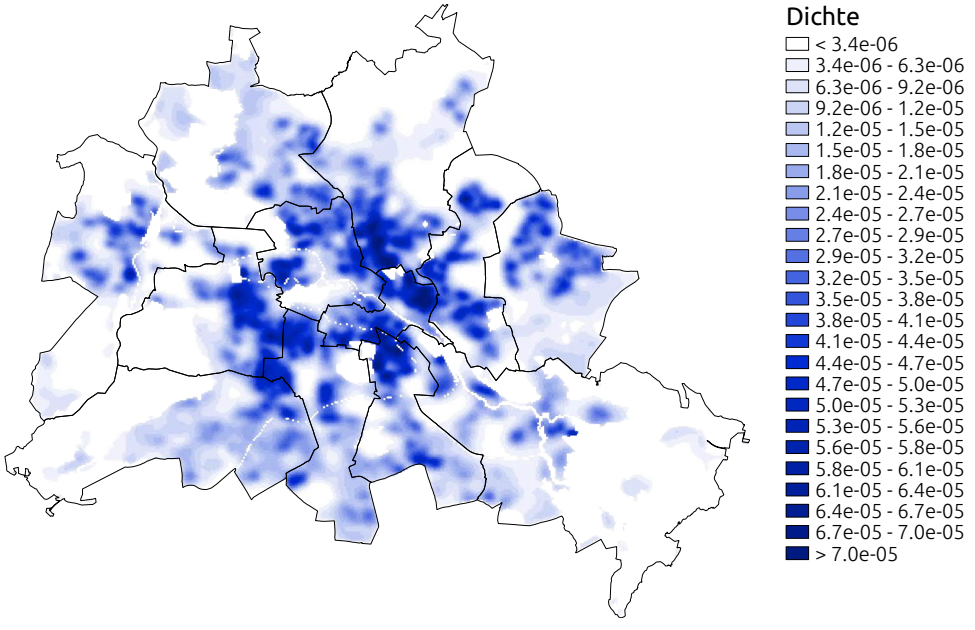
Zwei wesentliche Entscheidungskriterien innerhalb des Verfahrens sind die Anzahl an Iterationen *i* und die Länge der Burnin-Phase *b*, welche das Verfahren durchlaufen soll. Um den Einfluss dieser Parameter auf das Endergebnis abschätzen zu können, wurden die Schätzungen für verschiedene Iterationszahlen durchgeführt und verglichen. So kann die Differenz zur wahren Adressdichte abhängig von der Iterationszahl beziehungsweise von der Laufzeit des Verfahrens bestimmt werden.

Ein Ergebnis des Kernelheaping-Verfahrens, als weiteres punktbezogenes Schätzverfahren, visualisiert ↘ Grafik 6 ebenfalls für die Ebene der Urnenwahlbezirke. Es ist gut sichtbar, dass unter Verwendung des Kernelheaping-Verfahrens deutlich weniger Details verloren gehen als bei der naiven Kerndichteschätzung, insbesondere unter Verwendung der Plugin-Methode. Zudem gelingt es dem Verfahren besser, die flächenbezogene Inhomogenität der einzelnen Bezirke auszugleichen. Dadurch werden sowohl die kleinen Bezirke im Zentrum als auch die großen Flächen am Stadtrand gut geschätzt. Beispielsweise sticht das Märkische Viertel im Bezirk Reinickendorf im Norden Berlins als eine Art „Fragezeichen“ in der wahren Dichte auf Urnenwahlbezirksebene gut erkennbar hervor (siehe Grafik 1). In den naiven Kerndichtekarten ist dieses Detail nicht gleichermaßen gut ausgeprägt.

An dieser Stelle sei darauf hingewiesen, dass die bessere Detailwiedergabe ohne eine explizite Angabe einer Bandbreite erreicht wird. Dies ist ein wesentlicher Vorteil gegenüber der naiven Kerndichteschätzung. Zudem konnte in der Arbeit gezeigt werden, dass die Wahl der Anzahl an Iterationen nur eine geringe Auswirkung auf das Ergebnis hat.

Grafik 6

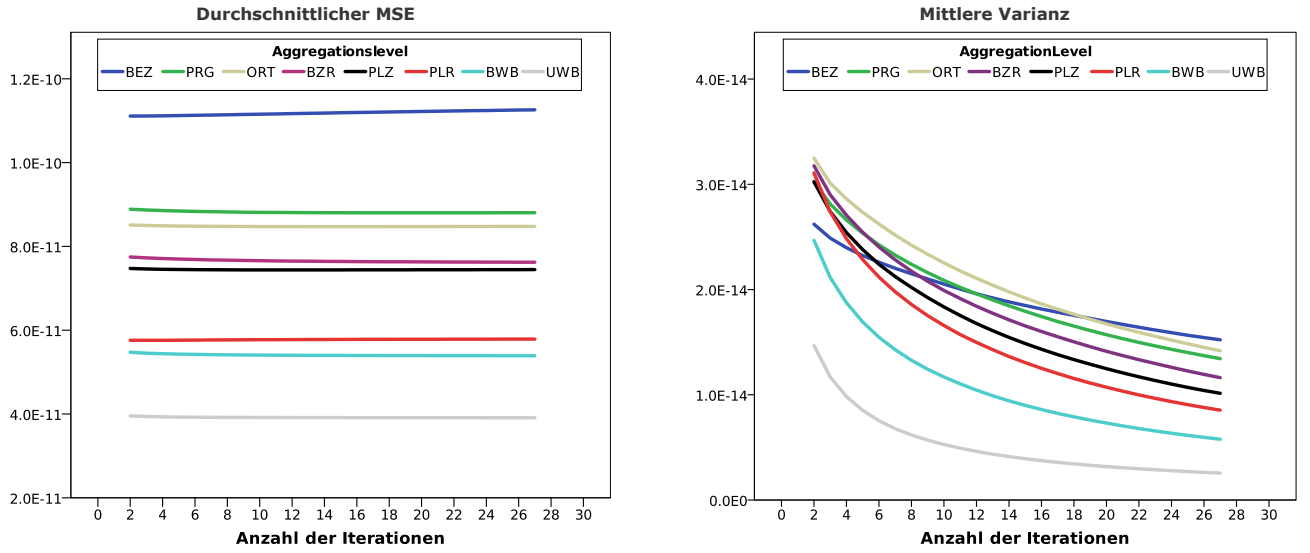
Kartografische Darstellung für die Kerndichteschätzung mit dem Kernelheaping-Verfahren



Dargestellt ist die Verteilung von Wahlberechtigten auf Urnenwahlbezirke in Berlin zum Stichtag 31.12.2016.

2020 - 01 - 0167

Grafik 7
Direkter Vergleich des mittleren quadratischen Fehlers (MSE) und der mittleren Varianz für das Kernelheaping-Verfahren, abhängig von der Iterationszahl für acht verschiedene Aggregationslevel



BEZ: Bezirk, PRG: Prognoseräume, ORT: Ortsteile, BZR: Bezirksregionen, PLZ: Postleitzahlen, PLR: Planungsräume, BWB: Briefwahlbezirke, UWB: Urnenwahlbezirke.

2020 - 01 - 0168

↘ Grafik 7 zeigt, dass mit zunehmender Iterationszahl der MSE weitgehend stabil bleibt, während gleichzeitig die Varianz abnimmt. Allerdings spielt die Varianz im Vergleich zu den Größenordnungen des Bias kaum eine Rolle.

8

Fazit

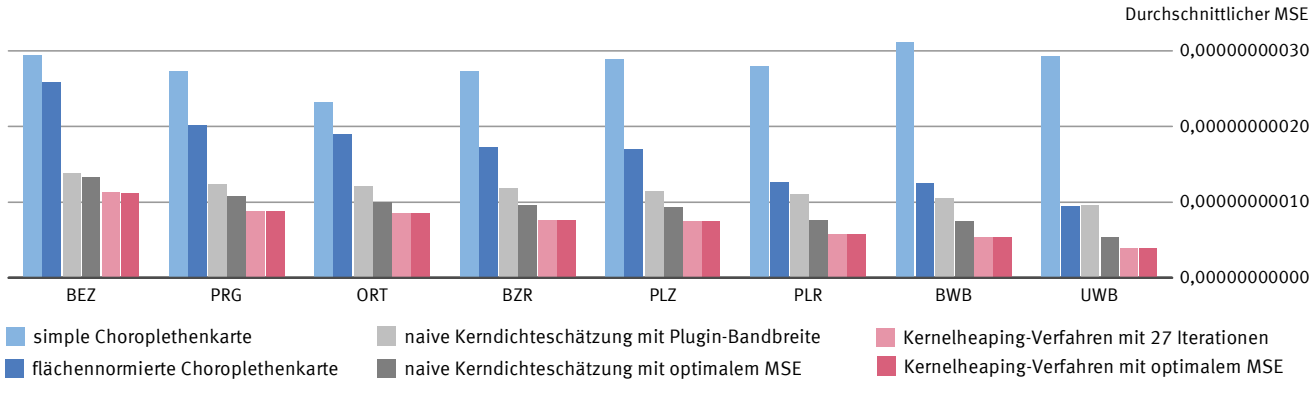
Wie oben bereits beschrieben, wird mit der Berechnung des MSE je Pixel (MSE-Karten) nur ein pixelweise gültiges Maß für die Qualität einer Schätzung wiedergegeben. Auf dieser Grundlage kann jedoch nicht die ganzheitliche Güte der Schätzergebnisse objektiv beurteilt werden. Daher wird für jede Karte der durchschnittliche MSE über alle Pixel berechnet. So ergibt sich ein globales Maß für die Qualität einer Schätzung, mit dessen Hilfe die Verfahren verglichen werden können.

Dieser durchschnittliche MSE über alle Pixel wird in ↘ Grafik 8 noch einmal – abhängig vom Aggregationslevel – für alle betrachteten Verfahren in einem Diagramm gegenübergestellt. Dabei sind die Verfahren farblich gruppiert. Da die Skala der Y-Achse auf null gesetzt ist, lässt sich das Verhältnis der MSE-Werte zwischen den Verfahren ablesen. Es zeigt sich beispielsweise, dass das Kernelheaping-Verfahren gegenüber der naiven Kerndichte auf Ebene der Urnenwahlbezirke einen um etwa 25 % geringeren durchschnittlichen MSE aufweist. Ebenfalls ersichtlich ist, dass die flächennormierte Choroplethenkarte auf der Aggregationsstufe der Urnenwahlbezirke nur etwa ein Drittel des MSE der simplen Choroplethenkarte besitzt.

Bei den einfachen Choroplethenkarten fällt auf, dass bei der simplen Schätzung eine genauere Aggregation nicht gleichermaßen zu einer besseren Schätzung führt. Intuitiv betrachtet, sollte sich mit detaillierteren Daten auch eine Besserung der Schätzergebnisse einstellen. Diese besitzen jedoch mit einer besseren Datenerhebung keinen messbaren Informationsgewinn. Die flächennormierte Choroplethenkarte hingegen weist solche Verbesserungen auf.

Für die naive Kerndichteschätzung werden mit optimaler Bandbreite zwar zunächst solide Ergebnisse erzielt, diese beruhen jedoch auf der Kenntnis der wahren Dichte, welche in der Praxis nicht zur Verfügung steht. Es ist ebenfalls erkennbar, dass sich mit einem höheren Detailgrad der Aggregationslevel auch ein geringerer durchschnittlicher MSE einstellt. Wird die naive Kerndichte ohne Kenntnis der wahren Dichte berechnet und

Grafik 8
Zusammenfassender Vergleich der mittleren quadratischen Fehler (MSE) aller betrachteten Verfahren für alle Aggregationslevel



BEZ: Bezirk, PRG: Prognoseräume, ORT: Ortsteile, BZR: Bezirksregionen, PLZ: Postleitzahlen, PLR: Planungsräume, BWB: Briefwahlbezirke, UWB: Urnenwahlbezirke.

2020 - 01 - 0169

ein Plugin-Bandbreitenselektor genutzt, können die Ergebnisse gegenüber dem Kernelheaping-Verfahren nicht überzeugen. Auf Ebene der Urnenwahlbezirke liegt die naive Kerndichte mit der flächennormierten Choroplethenkarte gleichauf.

Insgesamt lässt sich feststellen, dass das Kernelheaping-Verfahren über alle Aggregationsstufen hinweg das beste Ergebnis in Hinblick auf den minimalen durchschnittlichen MSE liefert. Die Iterationszahl spielt dabei eine untergeordnete Rolle für die Qualität des Schätzergebnisses. Demzufolge zeigt sich, dass das Kernelheaping-Verfahren auf Realdaten praxistauglich und robust einsetzbar ist.

Neben der quantitativen Bewertung ist auch optisch eine detailliertere Dichteschätzung erkennbar. Durch das punktbasierte Schätzverfahren werden einzelnen Pixeln des Bildes separate Dichtewerte zugeordnet. Dadurch gelingt es mit dem Kernelheaping-Verfahren, feine Strukturen in stark besiedelten Gebieten herauszuarbeiten, während gleichzeitig in großflächigen Arealen homogen geschätzt wird. Die einfache Kerndichteschätzung, als weiteres punktbasiertes Schätzverfahren, konnte hier keine gleichwertigen Ergebnisse erzielen, da mit einer fix gewählten Kernelgröße nicht adäquat auf die inhomogenen Aggregationen eingegangen werden kann. Die Choroplethenkarten geben die wahre Dichte am schlechtesten wieder, haben jedoch durch ihre Eigenschaft, insbesondere auf detaillierten Aggregationsleveln optisch ansprechende, scharfe Konturen zu bilden, ihre Vorteile.

Das Kernelheaping-Verfahren zeigt sich robust hinsichtlich der Parameterwahl. Auch wenn für die hier verwendete Adressdichte optimale Parameter ermittelt werden konnten, liefert das Kernelheaping auch bei vordefinierten, pauschalen Parametern ähnlich gute Resultate. Dies ist ein weiteres, wesentliches Unterscheidungsmerkmal gegenüber den einfachen Kerndichteschätzungen, welche mit den in dieser Arbeit verwendeten Plugin-Bandbreiten teilweise unbrauchbare Ergebnisse lieferten.

Ein Nachteil des Kernelheapings sind die hohen Ressourcenanforderungen. Sie sind im Vergleich zu den Choroplethen- und einfachen Kerndichtekarten sehr viel umfangreicher. Ein weiterer potenzieller Nachteil ist die nicht deterministische Berechnung. Durch den stochastischen Anteil entsteht eine gewisse Unsicherheit des Ergebnisses. Es konnte jedoch gezeigt werden, dass diese Varianz für die betrachteten Adressdaten im Vergleich zum Bias um den Faktor 10^3 geringer ist und damit in der Praxis keine Rolle spielt. Im Hinblick auf Geheimhaltungsaspekte könnte sich diese Unsicherheit wiederum als Vorteil herauskristallisieren. Damit stellt das Kernelheaping-Verfahren eine praxistaugliche Alternative für die Dichteschätzung für Wahldaten dar.

LITERATURVERZEICHNIS

Duong, Tarn. *Kernel Smoothing*. 2017. https://cran.r-project.org/web/packages/ks/index.html, R package Version 2.0.

Fahrmeir, Ludwig/Hamerle, Alfred/Tutz, Gerhard. *Multivariate statistische Verfahren*. 2. Auflage. Berlin/New York 1996.

Groß, Marcus/Rendtel, Ulrich. *Kernel Density Estimation for Heaped Data*. In: Journal of Survey Statistics and Methodology. Jahrgang 4. Ausgabe 3/2016, Seite 339 ff.

Groß, Marcus. *Kernelheaping: Kernel Density Estimation for Heaped and Rounded Data*. 2017. https://cran.r-project.org/web/packages/Kernelheaping/index.html, R package Version 2.0.

Groß, Marcus. *Messfehlermodelle für die Survey-Statistik und die Wirtschaftsarchäologie*. Dissertation. Freie Universität Berlin, 2016. [Zugriff am 9. März 2020.] Verfügbar unter: https://refubium.fu-berlin.de/handle/fub188/9385

Groß, Marcus/Rendtel, Ulrich/Schmid, Timo/Schmon, Sebastian /Tzavidis, Nikos. *Estimating the density of ethnic minorities and aged people in Berlin: multivariate kernel density estimation applied to sensitive georeferenced administrative data protected via measurement error*. In: Journal of the Royal Statistical Society: Series A (Statistics in Society). Jahrgang 180. Ausgabe 1/2017, Seite 161 ff.

Härdle, Wolfgang/Müller, Marlene/Sperlich, Stefan/Werwatz, Axel. *Nonparametric and Semiparametric Models*. Springer Series in Statistics. Berlin/Heidelberg 2004.

Silverman, Bernard W. *Density Estimation for Statistics and Data Analysis*. Monographs on Statistics and Applied Probability 26. London 1986.

Stadtportal, BerlinOnline. *Berlin Open Data*. 2017. Geometrien und Wahl- sowie Wahlstrukturdaten stehen unter der offenen Lizenz CC-BY und sind über das Open Data Portal des Landes Berlin verfügbar. [Zugriff am 9. März 2020.] Verfügbar unter: https://daten.berlin.de [10.10.2017]

Wand, Matt P./Jones, Chris. *Multivariate plug-in bandwidth selection*. In: Computational Statistics. Jahrgang 9. Ausgabe 2/1994, Seite 97 ff.