

Datenklassifikation: Mehrwert oder Manipulation?

Jochen Schiewe¹

¹ HafenCity Universität Hamburg, Labor für Geoinformatik und Geovisualisierung (g2lab)

DOI: 10.5281/zenodo.14616647

Zusammenfassung: Dieser Beitrag behandelt zwei grundlegende Probleme bei der Datenklassifikation, d. h. der Gruppierung quantitativer Werte, z. B. für die Erzeugung von Choroplethenkarten. Das erste Problem besteht darin, dass konventionelle Verfahren wie äquidistante oder Quantil-Einteilungen keine Bewahrung räumliche Muster in mono-temporalen sowie von Veränderungseigenschaften in multi-temporalen Datensätzen vorsehen. Es werden daher entsprechende Verfahren vorgestellt, die eine entsprechende Optimierung vornehmen (*Feature-Preserving Data Classification*). Das zweite Problem besteht darin, dass es weder Standardverfahren, noch „richtige“ oder „falsche“ Methoden für die Datenklassifikation gibt bzw. geben kann. Es wird ein zweckorientierter Einsatz verbunden mit einem kritischen Umgang und einer Bereitstellung von Metadaten vorgestellt.

Schlüsselwörter: Datenklassifikation, Choroplethenkarte, räumliche Muster, Veränderungen

1 Einführung

Bei der Verarbeitung von quantitativen Daten wird häufig eine Datenklassifikation vorgenommen. Eine solche Klassifikation dient z. B. der Reduktion von Datenmengen, Rechenzeiten bei Verarbeitungen und ganz besonders der Erzeugung einer kontrastreicher Visualisierung (z. B. in thematischen Karten, bei denen Klassen „bis zu 100 Einwohner pro km²“, usw. gebildet werden). Im Folgenden wird die Anwendung der Datenklassifikation für solche Choroplethenkarten aus Illustrationszwecken im Fokus stehen.

Ein Problem bei der Auswahl und Anwendung von Methoden der Datenklassifikation besteht darin, dass nach Anwendung der Verfahren

- in mono-temporalen Karten wichtige Informationen (insbesondere räumliche Muster wie Extremwerte, Cluster oder Hot Spots) sowie
- in multi-temporalen Karten (wiedergegeben als Kartenanimationen oder sog. *small multiples*) gewisse Veränderungseigenschaften (z. B. große Werteunterschiede oder Trends)

nicht mehr erkennbar sein können. Die unterschiedlichen GIS-Softwareprodukte bieten Standardverfahren und -einstellungen an (z. B. äquidistant, Quantile), die solche Eigenschaften per se nicht bewahren und häufig unreflektiert eingesetzt werden.

Darüber hinaus wird im übergeordneten Kontext das Thema der Datenklassifikation – z. B. auch im Feld der Kritischen Kartographie - kontrovers diskutiert, bis hin zum Vorwurf der Manipulationsmöglichkeit. Dies liegt zum einen daran, dass eine Klassifikation immer mit einem Informationsverlust verbunden ist. Zum anderen erzeugen verschiedene Methoden für ein und denselben Datensatz sehr unterschiedliche Ergebnisse, z. B. bei der Erzeugung von Choroplethenkarten, und daraus resultierende Interpretationen (Abschnitt 2).

Basierend auf diesen Problemstellungen bestehen die Ziele dieses Beitrages darin,

- alternative, erweiterte Methoden der Datenklassifikation vorzustellen, die eine Bewahrung räumlicher Muster (Abschnitt 3) bzw. zeitlicher Veränderungen (Abschnitt 4) bewahren können, und
- einen kritischen Umgang bei der Anwendung von Methoden der Datenklassifikation zu verfolgen (Abschnitt 5).

2 Datenklassifikation: Status Quo

2.1 Varianten

Eine Datenklassifikation bei der Gestaltung von Choroplethenkarten ist optional. Ein Verzicht auf die Klassifikation bedeutet jedoch eine größere visuelle Heterogenität, geringere visuelle Kontraste und eine schwierigere Extraktion von Werten aus den nicht klassifizierten Farbwerten.

Entscheidet man sich für eine Datenklassifikation, gibt es eine Reihe unterschiedlicher, auch in der Standard GIS-Software implementierter Verfahren, die unterschiedliche Regeln und Zwecke verfolgen. Ein gängiger und intuitiv „erwartbarer“ Ansatz ist die *äquidistante Gruppierung*. Diese gleichabständigen Klassengrenzen sind jedoch für schiefe Datenverteilungen weniger geeignet, da eine sehr homogene Kartenansicht erzeugt wird und viele Unterschiede nicht angezeigt werden können (Abb. 1, links).

Alternativ wird eine Gleichverteilung der Werte in Klassen und einem resultierenden, verbesserten ästhetischen Eindruck postuliert. (Gruppierung nach *Quantilen*; Abb. 1, mitte). Sollen einzelne Extremwerte farblich hervorgehoben werden, ist diese Einteilung allerdings ungeeignet, da die obersten und untersten Klassen i. d. R. mehr als einen Wert beinhalten.

Ein anderer Ansatz ist die Berücksichtigung von statistischen Parametern. Hier gibt es z. B. die Methode der *natural breaks* (auch: *Jenks optimal*), die Grenzen an lokalen Minima im Histogramm setzt (Abb. 1, rechts), oder die Histogramm-Angleichung (z. B. mit engen Klassenbreiten in der Mitte einer Normalverteilung).

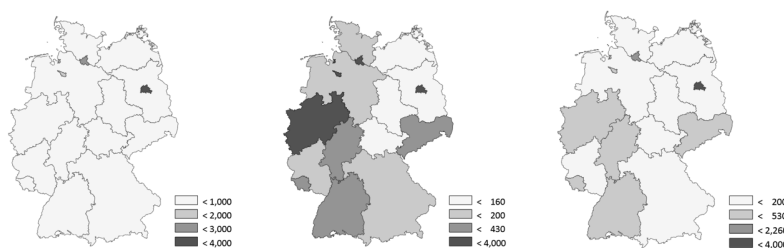


Abb. 1: Anwendung verschiedener Methoden zur Datenklassifikation für einen Datensatz (Bevölkerungsdichte in Deutschland, 1995), v.l.: äquidistant, Quantile, *natural breaks* (nach Schiewe 2022)

Aus den unterschiedlichen Zwecken wird bereits deutlich, dass es keine „Standard“-Methode geben kann. Abb. 1 verdeutlicht zudem, dass unterschiedliche Methoden sehr unterschiedliche Ergebnisse – und damit unterschiedliche Interpretationen – erzeugt werden.

Alle diese Methoden sind rein datenbasiert - räumliche Beziehungen (wie z. B. Nachbarschaften) und insbesondere die Erhaltung räumlicher Muster werden nicht berücksichtigt und

können bei der Klassifizierung verloren gehen. Andererseits ist gerade die Erkennung von räumlichen Mustern eine zentrale Aufgabe bei Choroplethenkarten, die durch die Klassifikation erschwert oder gar unmöglich gemacht werden kann.

2.2 Grundlegende Arbeiten

Das Thema Datenklassifikation für kartographische Zwecke wird z. B. in den Überblicksartikeln von Cromley & Cromley (1996) oder Coulsen (1987) umfangreich behandelt. Der Schwerpunkt liegt dabei auf datengetriebenen Methoden. Einige empirische Studien beschäftigen sich auch mit dem Vergleich solcher Methoden zur Beantwortung typischer Kartennutzungsaufgaben (z. B. Goldsberry & Battersby 2009; Brewer & Pickle 2002). Darüber hinaus wurden interaktive Tools entwickelt, um die „optimale“ Wahl für eine bestimmte Anwendung zu finden, z. B. durch die Verwendung von verknüpften Ansichten zwischen Datenhistogramm und Choroplethenkarte (Andrienko et al., 2001).

Über das bereits erwähnte Problem der Vernachlässigung des räumlichen Kontextes ist relativ wenig veröffentlicht worden; Armstrong et al. (2003) geben dazu einen Überblick. Versuche, Wertunterschiede von räumlichen Nachbarn zu beschreiben und zu erhalten, wurden z. B. von Smith (1986) oder Monmonier (1972) veröffentlicht. Häufig wird eine Vereinfachung der dargestellten Muster angestrebt, damit der Kartennutzer die groben Trends in der Darstellung schneller erfassen kann, ohne durch einen detaillierten und „fleckigen“ Eindruck gestört zu werden (Andrienko et al. 2002; MacEachren 1994). Chang & Schiewe (2019) haben ein Konzept zur Erhaltung ausgewählter räumlicher Muster vorgestellt – dieser Ansatz wird in diesem Beitrag wieder aufgegriffen (Abschnitt 3).

Ferner gibt es keine systematische Behandlung von multi-temporalen Datenklassifikationen in der Literatur. Schiewe (2023) hat ein Konzept erstellt und für die Erhaltung einfacher Veränderungseigenschaften eingesetzt. Cybulski (2022) führte eine empirische Studie zu den Auswirkungen zeitlicher Trends von Mustern in Choroplethenkarten durch und kam ebenfalls zu dem Schluss, dass in Zukunft ein erhöhter Bedarf an effektiver und effizienter kartographischer Visualisierung multitemporaler Daten besteht.

3 Bewahrung räumlicher Muster

Die Tatsache, dass die o. g. Standardverfahren der Datenklassifikation rein datenbasiert arbeiten und keine Nachbarschaften berücksichtigen, führt dazu, dass gewünschte räumliche Auffälligkeiten oder Muster nicht garantiert bewahrt werden können. Abb. 2 macht dies am Beispiel eines lokalen Maximums deutlich, das je nach Klassifikation als solches „graphisch ablesbar“ ist oder nicht.

Chang & Schiewe (2019) bzw. Schiewe (2021) haben eine Methode entwickelt, die ausgehend von der a priori Spezifikation eines möglichen Musters (z. B. lokales Extremum, Hot/Cold Spot, Cluster) eine optimale Bewahrung dieses Musters für einen gegebenen Datensatz ermittelt. Hierzu wurde auch ein entsprechendes QGIS-Plugin entwickelt (<https://gitlab.com/g2lab/aChor>). Die grundlegende Idee besteht darin, dass zwischen Originalwerten von möglichst allen Regionen, die einen signifikanten Unterschied aufweisen (z. B. lokales Maximum vs. kleineren Werte oder Cluster vs. Nicht-Cluster) auch ein Klassenunterschied festgelegt wird (und damit auch ein Farbunterschied in der kartographischen Darstellung).

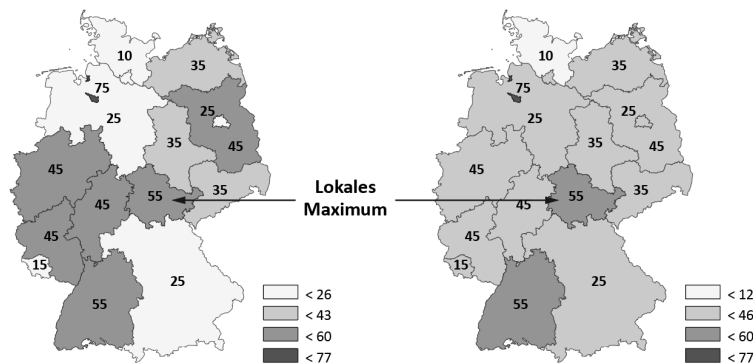


Abb. 2: Lokales Maximum (Werte größer als bei allen direkten Nachbarn) – nach Datenklassifikation nicht deutlich (da in selber Klasse wie einige Nachbarn; links) sowie bewahrt (rechts); Abbildung verändert nach Schiewe 2022

Hierzu werden alle betreffenden Wertepaare bestimmt und in einem *line sweep*-Diagramm aufgetragen. *Line sweep*-Verfahren haben sich bereits zu anderen Zwecken als ein effizientes Hilfsmittel bewährt, z. B. für eine räumliche Partionierung im Kontext einer Overlay-Operation (Kriegel et al 1992). Hier wird nun die orthogonale *line sweep* von links nach rechts über diese Intervalle bewegt. An allen Startwerten der Intervalle werden Haltepunkte definiert. Diese stellen eine mögliche, gewünschte Klassengrenze dar. Wenn c Klassen erforderlich sind, werden $c-1$ Linien erstellt. Für jede Linie an jedem Haltepunkt wird die Anzahl der Schnittpunkte gezählt. Dann werden alle Kombinationen aller Linien (für eine gegebene Anzahl von Klassen) in Betracht gezogen, um die beste Lösung zu finden (d. h. so viele Bedingungen wie möglich zu erfüllen).

Abb. 3 zeigt das Prinzip für die Aufgabe „Erhalte lokales Minimum“: Zuerst wird der nächstgrößere Wert in der Nachbarschaft des lokalen Minimums identifiziert und das entsprechende Intervall ($x = [51; 59]$) im *line sweep*-Diagramm aufgetragen. Dieses Vorgehen wird für alle anderen lokalen Minima im Datensatz wiederholt. Dann werden – für die vorgegebene Anzahl von drei Schnitten – die Platzierung der Linien gesucht, die möglichst alle Intervalle schneidet.

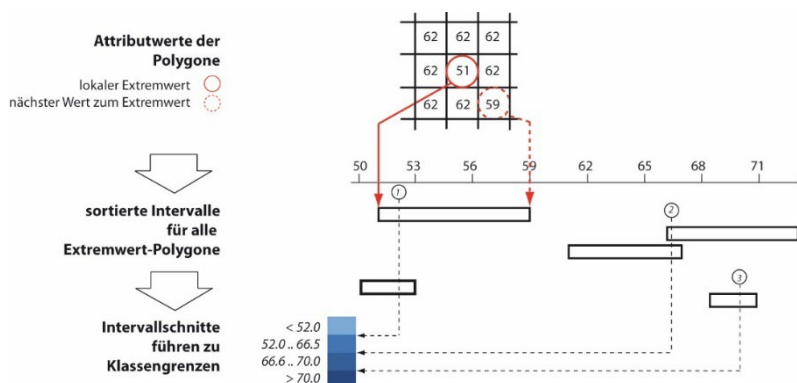


Abb. 3: Prinzip des Verfahrens zur Bewahrung lokaler Minima durch *line sweep*-Verfahren (*feature preserving data classification*)

Abhängig von der Anzahl und Verteilung der Intervalle sowie der Anzahl gewünschter Klassen kann es u. U. nicht möglich sein, alle gewünschten Intervalle zu schneiden. Das Verfahren kann also nur eine bestmögliche (d. h., optimale Lösung) finden. Im Vergleich zu konventionellen Verfahren (äquidistant, Quantile) konnte für diesen Ansatz aber eine stets bessere Bewahrung der räumlichen Muster nachgewiesen werden (Chang & Schiewe 2019).

4 Bewahrung Veränderungseigenschaften

Geht man von den oben verwendeten mono-temporalen Darstellungen auf multi-temporale Choroplethenkarten über, ergibt sich die Aufgabenstellung, bestimmte Veränderungen in der Zeitserie auch nach der Datenklassifikation zu bewahren, d. h. sichtbar zu machen.

Analog zum o. g. Vorgehen muss auch hier a priori definiert werden, welche Veränderungseigenschaften erhalten werden sollen. Dies ist allerdings komplexer als im mono-temporalen Fall - folgende Parameter gilt es zu betrachten:

- **Area of Interest:** Angabe der räumlichen Einheiten (z. B. Bundesländer) - von einer einzelnen Region bis hin zur Gesamtheit aller Regionen im Datensatz.
- **Time of Interest:** Angabe des Zeitintervalls (Start- und Endepoche) und der zeitlichen Auflösung (Lag).
- **Change Properties:** Diese können mit Hilfe verwandter (geo-)statistischer Maße beschrieben werden (in einigen Fällen durch Hinzufügen notwendiger Schwellenwerte):
 - Differenz der Attributwerte für ausgewählte räumliche Einheiten innerhalb des ausgewählten Zeitintervalls: globales Maximum/Minimum, lokales Maximum/Minimum, größte/kleinste Differenzen (z. B. ausgedrückt durch Perzentile) usw.
 - Trend: (monotoner) Anstieg/Abfall für ausgewählte räumliche Einheiten innerhalb eines ausgewählten Zeitintervalls, usw.
 - Multi-temporale, räumliche Muster: Zeitliche Veränderung der Kennwerte von Hot/Cold Spots, Clustern usw.
- **Visualisierungskontext:** Dieser berücksichtigt insbesondere die beabsichtigte Kartennutzung, den Kartennutzer, die Dauer der Präsentation oder das Kartenlayout. All diese Faktoren haben einen Einfluss auf die Anzahl von Klassen, aber auch auf mögliche Gestaltungsmerkmale für
 - die visuelle Betonung von Veränderungseigenschaften, oder
 - die Vermeidung von „False Positives“ (auch: „fake changes“), d. h., dass „unwichtige“ Veränderungen nicht in der gleichen Weise dargestellt werden wie „wichtige“ Veränderungen.

Basierend auf dieser a priori Festlegung werden nun wieder alle Paare oder Tupel von Originalwerten gesucht, die die Bedingung erfüllen (und erhalten bleiben sollen). Dieses Vorgehen wird in folgendem Beispiel demonstriert: Gesucht ist eine Datenklassifikation mit fünf Klassen, die die 90 % größten, positiven Attributwertänderungen für alle räumlichen Einheiten für alle gegebenen Zeitepochen mit einer zeitlichen Auflösung von einer Epoche beibehält, wobei diese Änderungen mit einer Klassendifferenz von 2 (zu Zwecken der visuellen Hervorhebung) dargestellt werden. Darüber hinaus sollen alle anderen Änderungen mit einer Klassendifferenz von weniger als 2 dargestellt werden, um keine „fake changes“ zu suggerieren.

Abb. 4 zeigt das resultierende *line sweep*-Diagramm und die erzeugten Klassengrenzen. In diesem Fall konnten alle Bedingungen erfüllt werden. Abb. 5 vergleicht eine konventionelle multi-temporale Choroplethenkarte basierend auf einer äquidistanten Einteilung (oben) mit der auf der Grundlage der *feature preserving data classification*.

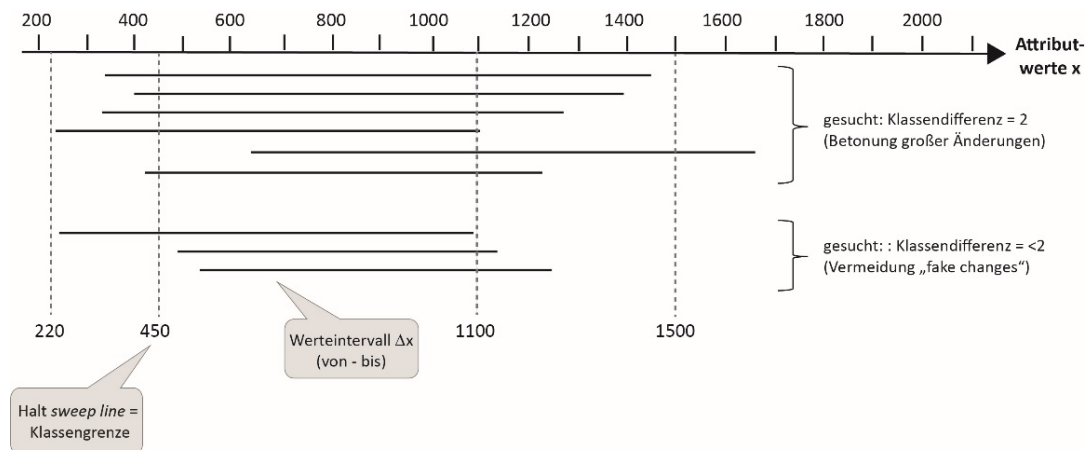


Abb. 4: line sweep-Verfahren für multi-temporalen Datensatz (Monatswerte Covid-19 für Januar bis Mai 2022; Daten: https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Daten/Inzidenz-Tabellen.html)

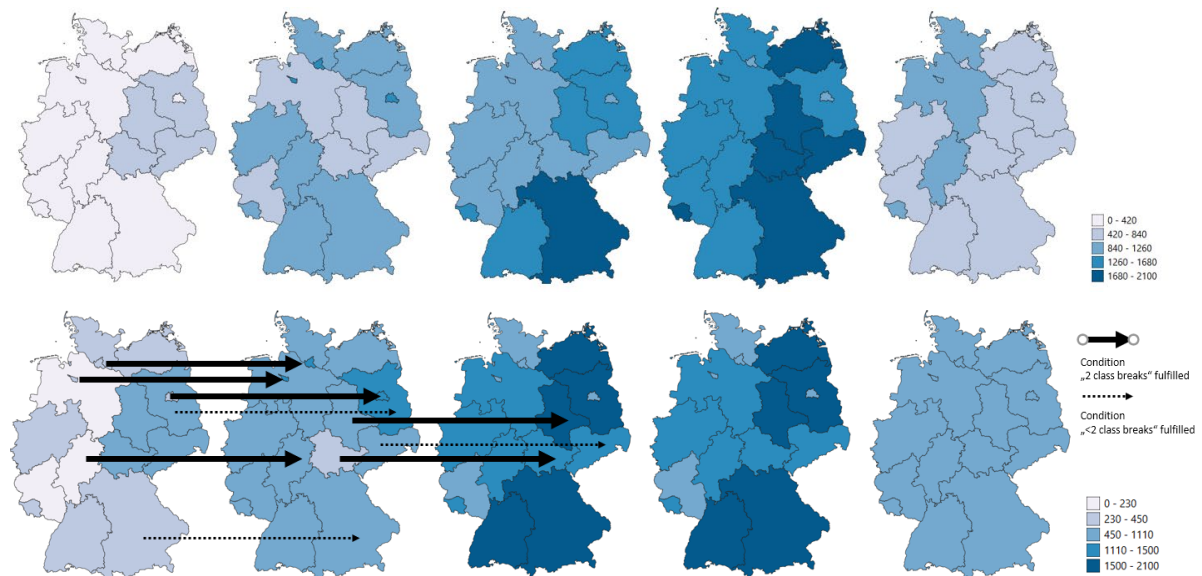


Abb. 5: Multi-temporale Choroplethenkarten für o. g. Datensatz Covid-19 – äquidistante Einteilung (oben) und *feature preserving data classification* (unten; mit Hervorhebung der Wertepaare, die die Veränderungsbedingungen erfüllt haben)

5 Mehrwert oder Manipulation?

Es wurde bereits ausgeführt, dass es keine „Standard“-Methode für die Datenklassifikation gibt bzw. geben kann, und andererseits sehr unterschiedliche Ergebnisse – sowie abgeleitete Interpretationen – erzeugt werden. Die in den Abschnitten 3 und 4 vorgestellten Verfahren zu Bewahrung räumlicher Muster sowie von Veränderungseigenschaften haben die Auswahlmöglichkeiten für die Durchführung der Datenklassifikation sogar noch erweitert.

Es bleibt festzuhalten, dass es keine „richtige“ oder „falsche“ Methode zur Datenklassifikation gibt (vorausgesetzt, die Methode selbst wird korrekt ausgeführt). Jedes einzelne Verfahren hat seine „Daseinsberechtigung“ – sei es, um übliche Erwartungen zu erfüllen (äquidistant), ein gutes Abbild der statistischen Eigenschaften zu geben (*natural breaks* oder Histogramm-Angleichung) oder aber ausgewählte räumliche Muster oder Veränderungseigenschaften zu bewahren (siehe Abschnitte 3 und 4). Aus dieser Vielfalt kann man aber durchaus auch einen Mehrwert ableiten: Es ist eine zielgerichtete Anpassung an einen Zweck zur Darstellung möglich. Dies setzt die a priori Definition eines solchen Zweckes voraus – was bei der Nutzung von Varianten oder gar Standardeinstellungen in der GIS-Software (z. B. *natural breaks* in QGIS bzw. ArcGIS) nicht geschieht. Und natürlich kann der Zweck auch – bewusst oder unbewusst – in einer Manipulation bestehen – und ist letztlich nicht auszuschließen.

Daraus ergibt sich ein Plädoyer für eine bewusste und kritische Anwendung der Datenklassifikation – sowohl von Seiten der Erstellenden, als auch von Seiten der Nutzenden. Dies beinhaltet insbesondere die Bereitstellung von Metadaten, die den Zweck verbal artikulieren und basierend auf verfügbaren Daten das Klassifikationsergebnis nachvollziehbar machen.

6 Zusammenfassung und Ausblick

Dieser Beitrag hat zwei grundsätzliche Probleme im Zusammenhang mit der Klassifikation von Daten (z. B. als vorbereitende Maßnahme zur Darstellung in Choroplethenkarten) aufgegriffen. Zum einen berücksichtigen konventionelle Verfahren (wie äquidistante oder Quantil-Gruppierungen) nicht die Bewahrung bestimmter räumlicher Muster (in mono-temporalen Karten) oder bestimmter Veränderungseigenschaften (in multi-temporalen Karten). Hierfür wurden Methoden vorgestellt, die basierend auf einer a priori Definition der zu bewahrenden Eigenschaften alle Wertepaare selektieren, die die Bedingungen erfüllen. Diese Paare werden in einem *line sweep*-Diagramm aufgetragen, in dem die Suche nach optimalen Klassengrenzen stattfindet.

Die geschilderte Vorgehensweise kann noch optimiert werden. Dies betrifft den hohen Rechenaufwand, der durch die Berücksichtigung aller Schnittkombinationen für mehrere *line sweep*-Linien entsteht – hier sind bereits angedachte Näherungslösungen weiter zu entwickeln. Ferner beruht die a priori Auswahl der zu bewahrenden Eigenschaften bisher auf einer subjektiven Vorauswahl. Hier ist es denkbar, dass eine vorgeschaltete, automatisierte Sequenz von geostatistischen Methoden alle relevanten Eigenschaften abprüfen kann.

Das zweite behandelte Problem betrifft die Tatsache, dass es kein Standardverfahren der Datenklassifikation geben kann, sondern eine Vielzahl von Varianten, die verschiedene Zwecke erfüllen können. Dies kann gleichermaßen als Mehrwert (flexibler Einsatz), als auch als Möglichkeit der Manipulation angesehen werden. Hieraus ergibt sich ein Plädoyer für eine bewusste und kritische Anwendung der Datenklassifikation. Dies sollte auch durch die Bereitstellung von Metadaten geschehen, die den Zweck verbal artikulieren und basierend auf verfügbaren Daten das Klassifikationsergebnis nachvollziehbar machen. Eine entsprechende Entwicklung von (OGC-) Standards in diesem Kontext ist wünschenswert.

Kontakt zum Autor:

Prof. Dr. Jochen Schiewe
HafenCity Universität Hamburg, Henning-Voscherau-Platz 1, 20457 Hamburg
Telefon: +49 40 428275442 / E-Mail: jochen.schiewe@hcu-hamburg.de

Literatur

- Andrienko, G., Andrienko, N. & Savinov, A.: Choropleth maps: Classification revisited. Proceedings International Cartographic Conference, 9 p., 2001
- Armstrong, M.P., Xiao, N. & Bennett, D.A.: Using Genetic Algorithms to Create Multicriteria Class Intervals for Choropleth Maps. *Annals of the Association of American Geographers*, 91(3): 595-623, 2003
- Brewer, C.A. & Pickle, L.: Evaluation of Methods for Classifying Epidemiological Data on Choropleth Maps in Series. *Annals of the Association of American Geographers*, 92(4): 662-681, 2002
- Chang, J. & Schiewe, J.: An open source tool for preserving local extreme values and hot/coldspots in choropleth maps. *KN – Journal of Cartography and Geographical Information*. 68(6): 307-309, 2019
- Cromley, E.K. & Cromley, R.G.: An analysis of alternative classification scheme for medical atlas mapping. *European Journal of Cancer*, 32A(9) 1551-1559, 1996
- Coulsen, M.R.C.: In the matter of class intervals for choropleth maps: With particular reference to the work of George Jenks. *Cartographica*, 24(2): 16-39, 1987
- Cybulski, P.: An Empirical Study on the Effects of Temporal Trends in Spatial Patterns on Animated Choropleth Maps. *ISPRS International Journal of Geo-Information* 11 (5) 273 DOI: 10.3390/ijgi11050273, 2022
- Goldsberry, K. & Battersby, S.: Issues of Change Detection in Animated Choropleth Maps. *Cartographica*, 44(3): 201-215, 2009
- Kriegel, H.-P., Brinkhoff, T. & Schneider, R. (1992): An Efficient Map Overlay Algorithm Based on Spatial Access Methods and Computational Geometry. In: Gambosi, G., Scholl, M., Six, H.-W. (eds.), "Geographic Database Management Systems", Workshop Proceedings, Capri, May 1991, Springer-Verlag: 194-211.
- MacEachren, A.M.: Some truth with maps: A primer on symbolization and design. Association of American Geographers, Washington, DC, 1994
- Monmonier, M.: Contiguity-biased class-interval selection and location allocation models. *Geographical Review*, 62: 203-228, 1972
- Schiewe, J.: Data classification methods for preserving spatial patterns, *Proc. Int. Cartogr. Assoc.*, 4, 95, <https://doi.org/10.5194/ica-proc-4-95-2021>, 2021.
- Schiewe, J.: *Kartographie. Visualisierung georäumlicher Daten*. Springer Spektrum. ISBN 978-3-662-65440-8, 2022
- Schiewe, J.: Preserving change information in multi-temporal choropleth maps through an extended data classification method. *The Cartographic Journal*, DOI: 10.1080/00087041.2023.2267944, 2023
- Smith, R.M.: Comparing traditional methods for selecting class intervals on choropleth maps. *The Professional Geographer*, 38(1): 62-67, 1986