Die Kerndichteschätzung als eine innovative Visualisierungsmethode georeferenzierter Daten in der amtlichen Statistik

Ein Methodenbericht

Swetlana Mamonova

Swetlana Mamonova M. Sc. ist Referentin im Referat „Informationsdienste, Regionalstatistik, Wahlen“ des Statistischen Landesamtes Baden-Württemberg.

In der amtlichen Statistik lassen sich kleinräumige georeferenzierte Daten mithilfe von INSPIRE Gitternetzen kartografisch gut darstellen. Allerdings sind aus Gründen der statistischen Geheimhaltungsrichtlinien Zellsperrungen notwendig, um Rückschlüsse auf Einzeldaten zu vermeiden. Das Geheimhaltungsverfahren der Zellsperrung geht allerdings mit einem Informationsverlust für kartografische Darstellungen einher. Das statistische Schätzverfahren der Kerndichteschätzung (Kernel Density Estimation – kurz KDE) hat gegenüber einer klassischen Datendarstellung mit Gitternetzen den Vorteil, dass ein hoher Informationsgehalt für die Gitterzellen sichergestellt wird und die statistische Geheimhaltung von Werten und Standorten der Einzeldaten gewahrt wird.

Die kartografische Darstellung von kleinräumigen georeferenzierten Daten stellt die amtliche Statistik, aus Gründen des Datenschutzes und der statistischen Geheimhaltung, vor besondere Herausforderungen. Sollen kleinräumige statistische Daten kartografisch dargestellt werden, so ist die Wahrscheinlichkeit erhöht, dass einzigartige Merkmalkombinationen auftreten, über die sich Rückschlüsse auf geheim zu haltende Fälle zurückführen ließen.[1] Um dem Anspruch der statistischen Geheimhaltung gerecht zu werden und andererseits die Nachfrage nach kleinräumigen Daten bedienen zu können, kann für die Visualisierung von kleinräumigen Daten auf die Kerndichteschätzung als ein mögliches Geheimhaltungsverfahren zurückgegriffen werden. Primär handelt es sich bei der Kerndichteschätzung um ein statistisches Verfahren zur Schätzung der Wahrscheinlichkeitsdichtefunktion einer Zufallsvariable aus einer Stichprobe. Es lässt sich aber zur Visualisierung und kartografischen Darstellung von kleinräumigen georeferenzierten Daten einsetzen, womit gleichzeitig die statistische Geheimhaltung von Einzeldaten gewährleistet wird.[2]

Methodenerläuterung

Das Verfahren der Kerndichteschätzung wird in der Statistik verwendet, um die Wahrscheinlichkeitsdichtefunktion einer Zufallsvariable aus einer Stichprobe zu schätzen. Es wird in der explorativen Datenanalyse und bei der Visualisierung von Daten eingesetzt, um ein besseres Verständnis über die Verteilung von Daten zu gewinnen. Bei der Kerndichteschätzung wird eine geschätzte Wahrscheinlichkeitsdichte erzeugt, indem um jeden Datenpunkt eine Kernfunktion gelegt wird und diese Kerne anschließend über den gesamten Datenraum aufsummiert werden. Die aufsummierte Dichte wird anschließend geglättet und liefert eine kontinuierliche Schätzung der Verteilungsfunktion. Die ermittelte Dichte gibt an, wie wahrscheinlich es ist, dass ein zufällig ausgewählter Punkt an einer Position in dem Untersuchungsraum erscheint.

Formel zur Berechnung der Kerndichteschätzung an einem Punkt x:

$$\hat{f}(x;h) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

K ist die Kernfunktion,

h die Bandbreite,

n die Anzahl der Datenpunkte.[3]

Parameter der Kerndichteschätzung

Für die Berechnung der Kerndichteschätzung gibt es eine Reihe von Parametern, die einen Einfluss auf das Ergebnis haben können.

Wahl der Kernfunktion

Ein entscheidender Parameter der Kerndichteschätzung ist die Wahl einer (symmetrischen) Kernfunktion, welche die Genauigkeit der Kerndichteschätzung maßgeblich beeinflusst. Gängige Kernfunktionen sind etwa die Gauß‘schen Kernfunktion oder die Epanechnikov-Kernfunktion, sie unterscheiden sich hinsichtlich ihrer Form und Ausdehnung. Je nach vorliegendem Datensatz können unterschiedliche Kernfunktionen sinnvoll sein.[4]

1 Große, Franziska/Schulz, Julian (2022): Verbleib und Herkunft von Pflegebedürftigen in Pflegeheimen in Niedersachsen 2019, Statistische Monatshefte 1/2022, Landesamt für Statistik Niedersachsen, S. 7.

2 IT.NRW, *https://www.it.nrw/kerndichteschaetzer-zur-veroeffentlichung-von-karten-mit-georeferenzierten-daten-der-amtlichen-0* (Abruf: 07.04.2024).

3 Silverman, Bernhard Walter (2003): Density Estimation for Statistics and Data Analysis. Chapman and Hall/CRC, S. 4–5.

4 Węglarczyk, Stanisław (2018): Kernel density estimation and its application, ITM Web of Conferences 23, 00037, XLVIII Seminar of Applied Mathematics, S. 2.

Bei der Kerndichteschätzung wird eine symmetrische Kernfunktion über einen Datenpunkt gelegt, wobei der Volumenwert unterhalb der Funktion eins beträgt. Hierbei ist der Dichtewert über der unmittelbaren Position eines Datenpunktes am höchsten und verringert sich mit zunehmender Entfernung vom Datenpunkt. Je mittiger ein Datenpunkt in einer Gitterzelle liegt, desto mehr trägt er zum Volumenwert der Gitterzelle bei, in der er verortet ist. Anschließend wird jeder Gitterzelle ein aufsummierter Dichtewert zugeordnet. Die Kerndichte ist somit auch von den Datenpunkten in den benachbarten Zellen abhängig.[5]

Größe der Bandbreite

Ein weiterer wichtiger Parameter der Kerndichteschätzung ist die Wahl einer angemessenen Bandbreite, die angibt, wie breit die Kernfunktion sein soll und eine Glättung der geschätzten Kernfunktionen leistet. Eine große Bandbreite führt zu einer Überglättung der geschätzten Dichte und einem generalisierten Dichte-Raster, während eine kleine Bandbreite ein detailliertes Raster erzeugt *(siehe Abbildung 1)*.[6] Die Wahl einer Bandbreite hat einen großen Einfluss auf die Genauigkeit der Schätzung der Wahrscheinlichkeitsdichte. Daher empfiehlt es sich gegebenenfalls mit der Größe der Bandbreite zu experimentieren, um die optimale Bandbreite für den vorliegenden Datensatz auszuwählen.[7]

Zellgröße

Die Wahl der Zellgröße beeinflusst die Dichtewerte der Kerndichteschätzung. Die kartografische Darstellung unterscheidet sich je nach Zellgrößen. Kleinere Zellgrößen erzeugen eine fließende, kontinuierlich wirkende Oberfläche. Größere Gitterzellen hingegen erzeugen ein grobkörniges Gesamtbild, die räumliche Verteilungsmuster der Daten verschwinden lassen können.[8]

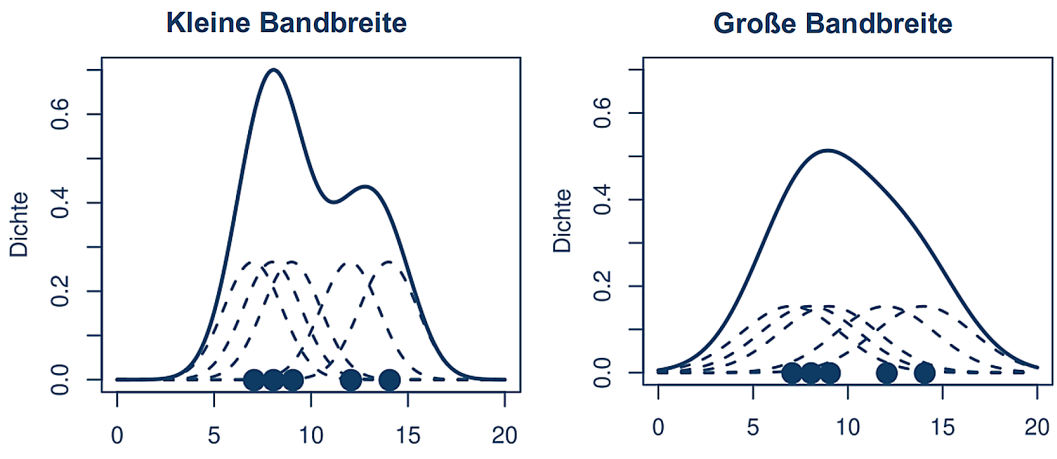
Klassifikationsverfahren und Klassenanzahl

Für die visuelle kartografische Darstellung der Kerndichteschätzung ist die Wahl des Klassifikationsverfahrens sowie die Anzahl der Klassen maßgeblich. Eine höhere Klassenanzahl ermöglicht zwar eine genauere Darstellung des Sachverhalts, lässt möglicherweise aber Rückschlüsse auf die Standorte geheim zu haltender Einzeldaten zu. Im Sinne der statistischen Geheimhaltung ist demnach eine geringe Klassenanzahl, für eine visuelle Darstellung, von Vorteil.

Interpretation der Kerndichtewerte

Bei der Interpretation der klassifizierten Kerndichtewerte gilt es zu beachten, dass diese nicht die exakte Anzahl der beobachteten Fälle bzw. deren georeferenzierten Standort widergeben. Die Werte der Kerndichteschätzung sind nicht ganzzahlig. Die kategorisierten Dichtewerte geben darüber Auskunft, wie viele Datenpunkte sich in unmittelbarer Nähe der Gitterzelle befinden. Im Allgemeinen weisen höhere Kerndichtewerte darauf hin, dass in diesem Bereich der Daten mehr Fälle vorliegen, während niedrigere Werte darauf hindeuten, dass in diesem Bereich weniger Fälle vorhanden sind.[9]

A1 Die Größe der Bandbreite beeinflusst das Ergebnis



5 IT.NRW, *https://www.it.nrw/kerndichteschaetzer-zur-veroeffentlichung-von-karten-mit-georeferenzierten-daten-der-amtlichen-0* (Abruf: 07.04.2024).

6 Quelle: Schärpler, Jörg-Peter: Kerndichteschätzungen im sozialräumlichen Bildungsmonitoring, *https://www.kommunales-bildungsmonitoring.de/fileadmin/user_upload/aktuelles/fachkonferenzen/2020_Materialien_Online-Dossier/Folien/Fachkonferenz_Bildungsmonitoring_2020_Fachvortrag_2_Schräpler.pdf* (Abruf: 04.01.2024).

7 IT.NRW, *https://www.it.nrw/kerndichteschaetzer-zur-veroeffentlichung-von-karten-mit-georeferenzierten-daten-der-amtlichen-0* (Abruf: 07.04.2024).

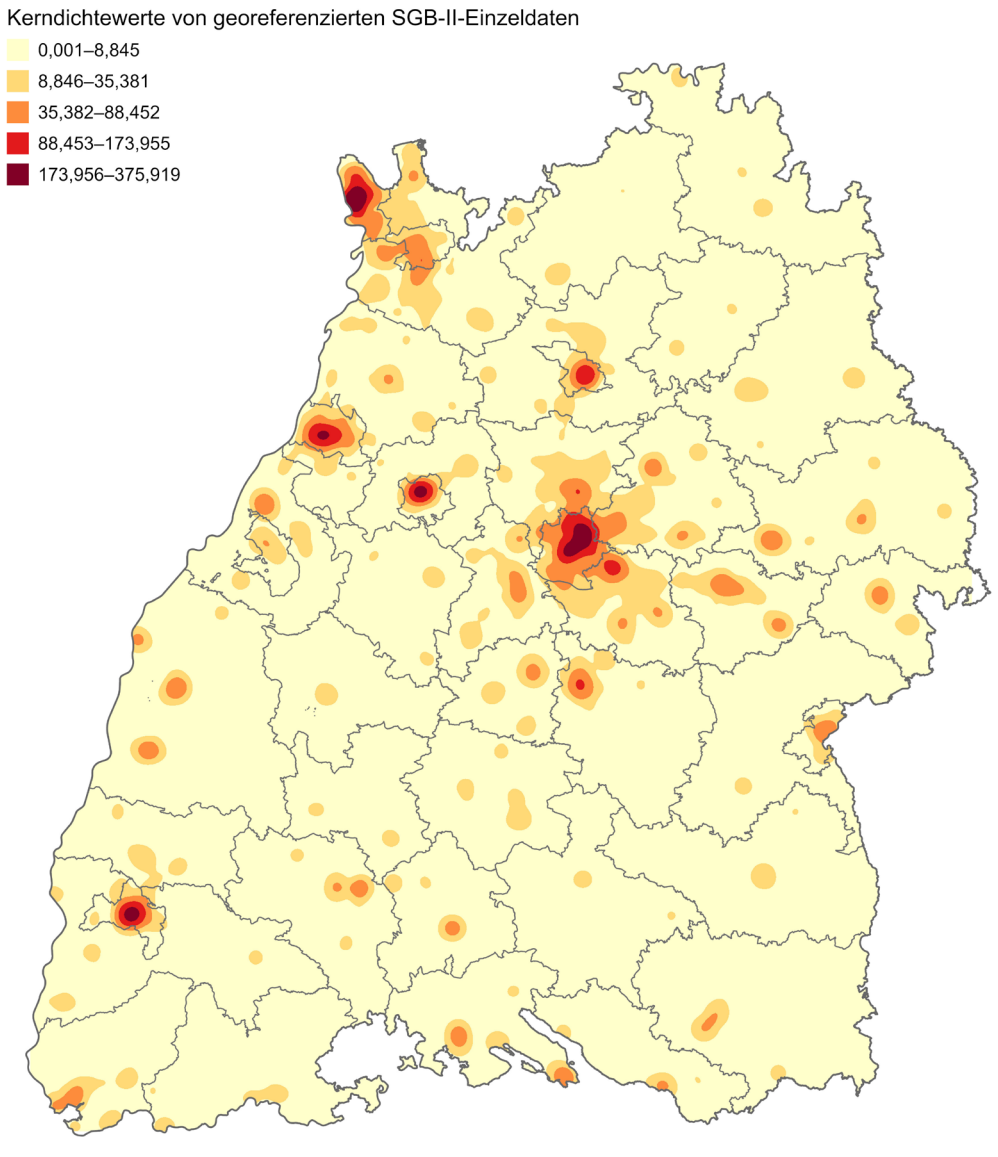
8 Gonschorek, Julia (2016): Zivile Sicherheit in urbanen Räumen – Adaption des KDE-Verfahrens zur optimierten Hotspot-Analyse für Behörden und Organisationen mit Sicherheitsaufgaben, in: AGIT – Journal für Angewandte Geoinformatik.

9 IT.NRW, *https://www.it.nrw/kerndichteschaetzer-zur-veroeffentlichung-von-karten-mit-georeferenzierten-daten-der-amtlichen-0* (Abruf: 07.04.2024).

Die Kerndichteschätzung am Beispiel eines geokodierten SGB-II-Adressdatensatzes

Im Folgenden soll die kartografische Visualisierung mithilfe einer Kerndichteschätzung anhand eines geokodierten, pseudonymisierten SGB-II-Adressdatensatzes (n=191 515) der Bundesagentur für Arbeit (BA) veranschaulicht werden *(Abbildung 2)*. Für diese Datensätze gelten Geheimhaltungsregeln, die eine kartografische Darstellung der Einzeldaten nicht zulässt, da sonst gegebenenfalls Rückschlüsse personenbezogener Daten anhand der genauen Standorte möglich wären. Mithilfe der Kerndichteschätzung lässt sich eine visuelle und kartografische Darstellung der adressgenauen Einzeldatensätze, unter der Einhaltung der statistischen Geheimhaltung, umsetzen. Die durchgeführte Kerndichteschätzung erfolgte mit der GIS-Software ArcGIS Pro, welche in der amtlichen Statistik eine Verbundstandard-GIS-Software ist. Die Berechnung der Kerndichteschätzung des SGB-II-Datensatzes erfolgte auf Grundlage eines 100 m x 100 m großen Gitters für Baden-Württemberg. Als Kernfunktion wurde der bei ArcGIS als standardmäßig gesetzte Epanechnikov-Kern herangezogen. Die Bandbreite wurde auf 4 Kilometer festgesetzt, da sie sich für eine großflächige, landesweite Betrachtung der Adress-

A2 Visualisierung von SGBII-Einzeldatensätzen in Baden-Württemberg 2021 mithilfe der Kerndichteschätzung



Datenquellen: SGB-II-Einzeldatensatz der Bundesagentur für Arbeit (2021).
Statistisches Landesamt Baden-Württemberg Landesinformationssystem.

© Kartengrundlage LGL, www.lgl-bw.de
Karte erstellt mit ArcGIS Pro 2024

daten als besonders geeignet erwiesen hat. Die Klasseneinteilung erfolgte nach der von der ArcGIS-Software vorgeschlagenen Methodik „Natürliche Unterbrechungen (Jenks)". Hierbei werden nummerische Werte einer Rangfolge von Daten überprüft, um ungleiche Verteilungen zu berücksichtigen. Dadurch wird eine ungleiche Klassenbreite mit unterschiedlichen Häufigkeiten von Beobachtungen pro Klasse erzielt. Die Farbkodierung der Klassen impliziert eine zunehmende Dichte der Datenpunkte (von gelb = niedrige Dichte bis rot = hohe Dichte). Eine standortgenaue Rückverfolgung der Datensätze ist nach der Kerndichteschätzung nicht mehr möglich. Die visuelle Darstellung der Kerndichteschätzung liefert ein gesamtheitliches Lagebild zu den räumlichen und regionalen Agglomerationen der SGB-II-Adressdateien in Baden-Württemberg.

Die Kerndichteschätzung in der kleinräumigen Betrachtung

Um den Mehrwert der Kerndichteschätzung für kleinräumige Daten anhand von Beispielen zu veranschaulichen, soll im folgenden anhand eines nicht näher definierten Raumausschnitts mit einem 100 m x 100 m großen Gitternetz das Problem der Zellsperrung veranschaulicht werden.

In *Abbildung 3* sind die Einzelwerte der SGB-II-Empfängerinnen und -Empfänger pro Gitterzelle zu entnehmen. Allerdings darf diese Darstellung für den betreffenden Datensatz aus Gründen der statistischen Geheimhaltung nicht veröffentlicht werden.

In *Abbildung 4* wurde die Zellsperrung für die SGB-II-Einzelwerte eins bis vier durchgeführt, allerdings hat dies zur Folge, dass die Karte mit einem Informationsverlust einhergeht. Räumliche Zusammenhänge sind in dieser Darstellung kaum bis gar nicht ersichtlich.

In *Abbildung 5* wurde die Kerndichteschätzung mit einer Bandbreite von 250 Meter durchgeführt. Die Darstellung der Einzeldatensätze schließt Rückschlüsse auf Einzelwerte und punktuelle Ereignisse aus, da eine ortsgenaue Zuweisung der Datenpunkte nicht mehr möglich ist. In dieser Darstellung zeichnen sich – im Gegensatz zu der Darstellung in *Abbildung 4* – räumliche Bezüge und ein Muster mit unterschiedlichen Punktdichten ab.

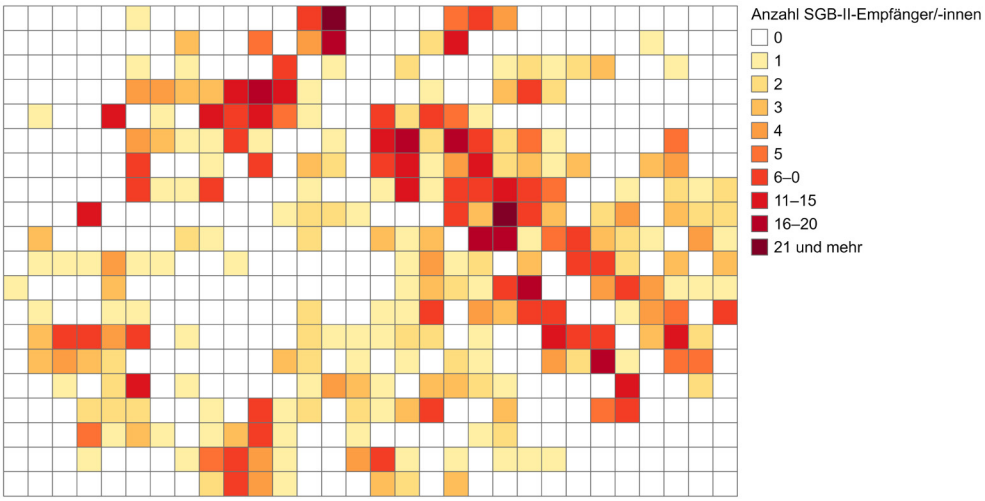
Auswirkungen von Barrieren auf die Dichteberechnung

In die Berechnung der Kerndichteschätzung lassen sich auch Barrieren, wie etwa das Straßennetz oder administrative Grenzen oder Geländereliefe einbinden. Somit können je nach Anwendungsfall und Fragestellung genauere Ergebnisse erzielt werden als bei der Berechnung ohne Barrieren. Dabei wird die Entfernung zum Beispiel entlang des Straßennetzes gemessen anstatt der kürzesten Entfernung zwischen Orten (Luftlinie).

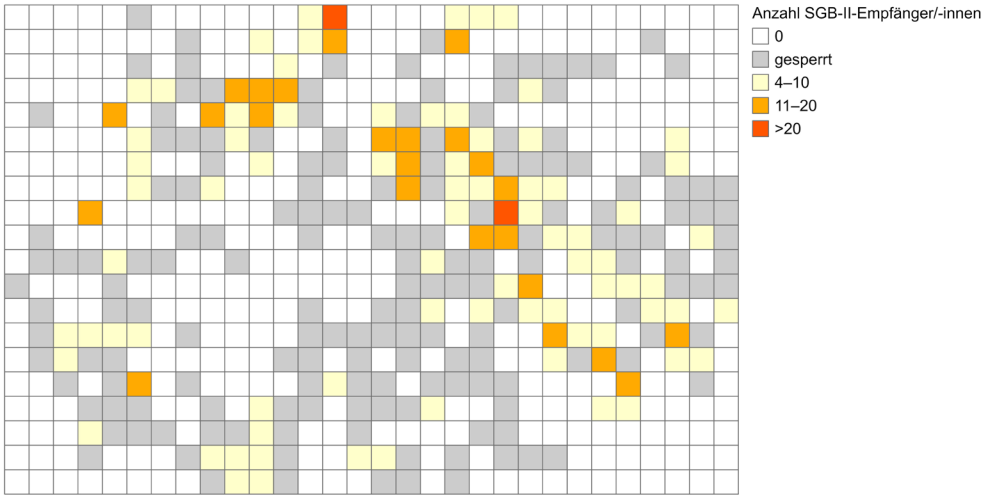
Fazit

Die Kerndichteschätzung ist ein geeignetes statistisches Schätzverfahren um kleinräumi-

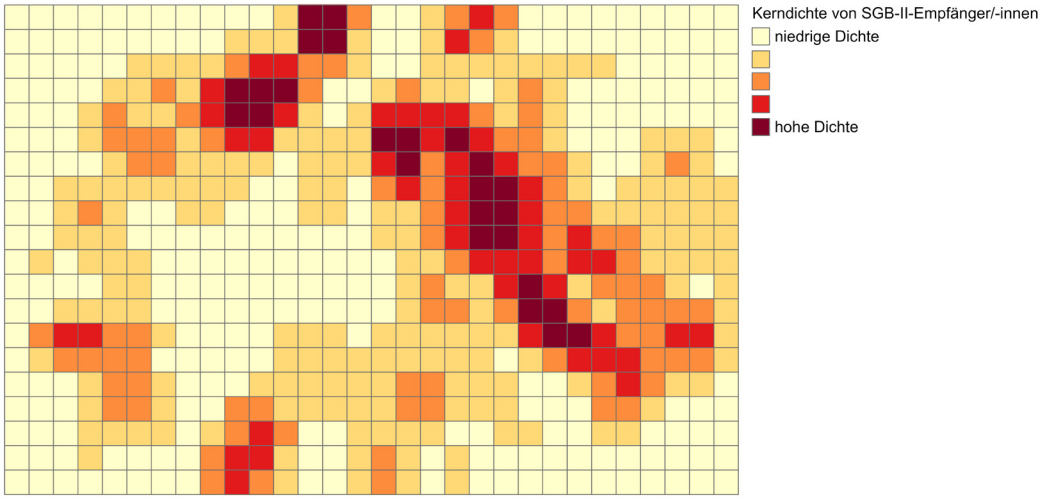
A3 Kartenausschnitt mit Einzelwerten



A4 Karte nach Zellsperrung



A5 Karte nach der Kerndichteschätzung



ge georeferenzierte Einzeldatensätze kartografisch darzustellen und zu veröffentlichen, während gleichzeitig der Datenschutz und die Geheimhaltung der Einzeldaten gewahrt wird. Die kartografische Visualisierung mithilfe der Kerndichteschätzung lässt Beobachtung und Rückschlüsse auf die räumlichen und regionalen Lagemuster und Konzentrationen der Datenpunkte zu. Die Kerndichteschätzung eignet sich insbesondere für georeferenzierbare Punktdaten wie beispielsweise Daten zu Verkehrsunfällen, zur Kriminalitäts- und Unternehmensstatistik, die Lohn- und Einkommensteuerstatistik oder die Statistik über beantragte Insolvenzverfahren. Durch die Berücksichtigung aller Einzeldaten im Schätzverfahren ergibt sich ein ganzheitliches räumliches Lagebild des entsprechenden Datensatzes mit einem höheren Informationsgehalt, als im Vergleich zu Kartendarstellungen, in denen Gitterzellen mit Einzelwerten, aus Gründen der Geheimhaltung, für eine Darstellung gesperrt werden müssen.

***www.statistik-bw.de**/SRDB/?E=GS*
Direkt zu ...
Regionaldaten

Weitere Auskünfte erteilt
Swetlana Mamonova,
Telefon 0711/641-28 46,
Swetlana.Mamonova@stala.bwl.de