

**Eine Methode zur Aufbereitung von georeferenzierten Punktdaten für eine in Bezug auf die Anonymisierungsproblematik unbedenklichen Nutzung durch externe Wissenschaftler über die kontrollierte Datenfernverarbeitung (kDFV) der Forschungsdatenzentren der statistischen Ämter (FDZ) oder der Bundesagentur für Arbeit (BA) – ein Kerndichteansatz**

Weisen Einzeldaten raumbezogene Referenzinformationen auf, handelt es sich in der Regel um Daten, die aufgrund der Anonymisierungsproblematik interessierten Wissenschaftlern nicht zur Verfügung gestellt werden können. So können z.B. pseudonymisierte Daten mit Adressangaben wie etwa die SGB II – Empfängerdatei der BA oder andere georeferenzierte Punktdaten aus Datenschutzgründen nicht an Wissenschaftler weitergegeben werden.

Diese Raumdaten stellen allerdings für die Wissenschaft eine sehr wertvolle Ressource dar. So bilden z.B. die Daten von SGB II- Empfänger bzw. deren Verteilung eine wesentliche Grundlage für Sozialstruktur- und Ungleichheits- sowie Segregationsanalysen. Von Interesse sind hier jedoch weniger die aus datenschutzrechtlichen Gründen mit Recht als problematisch anzusehenden Einzelfälle sondern i.d.R. die sich ergebenden Punktmuster bzw. „Dichten“ der vorliegenden Adressen oder georeferenzierten Punktdaten. Um diese sensiblen Daten trotzdem verarbeiten und nutzen zu können werden in einigen Bundesländern Quoten auf Gemeindeebene berechnet und der Wissenschaft zur Verfügung gestellt. Diese Daten sind jedoch für die meisten Analysen vollkommen ungeeignet:

1. Aufgrund der großen Heterogenität innerhalb der einzelnen Gemeinden ist die Aussagekraft der berechneten Quoten zweifelhaft, man erhält nicht informative gemittelte Quoten.
2. Die Verläufe der Häufigkeitsdichten sind i.d.R. nicht konform mit den administrativen Abgrenzungen diese Aggregatebenen. Einzelne Kommunen veröffentlichen zwar auch Quoten auf Stadtbezirksebene, aber selbst diese Aggregatebene ist z.B. in der Bildungsforschung für die statistische Beschreibung von Schuleinzugsgebieten viel zu ungenau und unbrauchbar.

In einigen Fällen, wie etwa der SGB II-Empfängerdatei liegen die Adressdaten zentral bundesweit vor und können mit einem geeigneten statistischen Verfahren so aggregiert und aufbereitet werden, dass einerseits landes- oder bundesweit sinnvolle Analysen auf kleinräumiger Ebene möglich sind und andererseits die Anonymität einzelner gewahrt wird bzw. keine Anonymisierungsproblematik entsteht.

Es kann gezeigt werden, dass mit Hilfe eines Kerndichteansatzes aus geographischen Koordinaten (wie etwa Adressdaten von SGB II-Empfängern) Häufigkeitsdichten und Dichteflächen erzeugt werden können, die im Hinblick auf die Anonymisierung vollkommen unbedenklich und unabhängig von vorgegebenen Gemeinde- oder Kreisabgrenzungen sind. Dieses Verfahren wurde in NRW inzwischen erfolgreich im Rahmen der Konstruktion von Sozialindizes für Schulen eingesetzt und wird von dem Ministerium für Schule und Weiterbildung zur Auswertung von Lernstandsergebnissen genutzt (siehe hierzu Schräpler 2011). Die Berechnungen erfolgten im Statistischen Landesamt NRW (IT.NRW).

Das besondere an dieses Verfahren ist, dass es auch dazu genutzt werden kann, externen nichtamtlichen Wissenschaftlern über die Forschungsdatenzentren der Statistischen Landesämter (FDZ) oder auch der Bundesagentur für Arbeit (BA) mittels kontrollierter Datenfernverarbeitung (kDFV) einen Zugang zu sinnvoll aggregierten georeferenzierten Punktdaten (Koordinaten) zu ermöglichen. Bei einem entsprechenden Vorgehen, wird der externe Wissenschaftler nie direkt mit georeferenzierte Einzeldaten in Kontakt kommen und als Ergebnis nur Flächendaten erhalten.

Ein wichtiger Aspekt bei der Anwendung dieses Verfahrens ist die Festlegung der Bandbreite des Kernschätzers. Je größer diese ist, desto stärker werden Häufigkeitsdichten geglättet, kleinere Bandbreiten offenbaren mehr Details. Der Forscher kann eine entsprechende Software (z.B. mit dem frei erhältlichen Statistikprogramm R) nutzen, um ein Programm zu schreiben, welches die optimale Bandbreite bei der Kerndichteschätzung ermitteln soll. Die konkrete Berechnung der Bandbreite erfolgt (u.U. über kDFV) vor Ort im FDZ an dem Originalmaterial.

Im FDZ werden dann mit Hilfe eines Kerndichteschätzers auf Basis der Bandbreite so genannte Shapefiles mit entsprechenden Dichteflächen bzw. Polygonen erzeugt. Die erzeugten Dichtekarten enthalten keine individuellen georeferenzierten Fälle mehr sondern nur noch Flächen, die die Punktdichte repräsentieren. Eine Reidentifizierung Einzelner ist normalerweise vollkommen ausgeschlossen (für NRW ergab sich z.B. eine optimale Bandbreite von 500m), dennoch sollte vor Weitergabe der Karte eine Prüfung im FDZ erfolgen. In Einzelfällen könnte die Bandbreite vergrößert werden, oder vorab eine minimale Bandbreite vorgeschrieben werden.

Für die praktische Durchführung müssen allerdings die Rechner der FDZ der Statistischen Ämter oder der BA mit Geoinformationssystemen ausgestattet und u.U. für Fernrechnen (Ausführen von Auswertungsprogrammen) vorbereitet werden. Zudem muss das Betreuungspersonal mit GIS-Programmen vertraut sein.

(Jörg-Peter Schräpler)

#### **Literatur:**

Schräpler, Jörg-Peter (2011): Konstruktion von SGB II-Dichten als Raumindikator und ihre Verwendung als Indikator im Rahmen der Sozialberichterstattung am Beispiel der "sozialen Belastung" von Schulstandorten - ein Kerndichte Ansatz. *ASTA Wirtschafts- und Sozialstatistisches Archiv*. Vol. 5, Nr. 2, S. 97-124.