Geheimhaltung

Statistische Geheimhaltung bei der Auswertung georeferenzierter Daten

von **Jörg Höhne** und **Julia Höninger**

Auch bei der Auswertung von georeferenzierten Statistiken, die dem Bundesstatistikgesetz oder Landesstatistikgesetzen unterliegen, muss die amtliche Statistik sicherstellen, dass das Statistikgeheimnis gewahrt wird. Bei der Anwendung der gängigen Geheimhaltungsregeln in der statistischen Geheimhaltungsprüfung ist die Mindestfallzahl bei georeferenzierten Daten ein besonderer Diskussionspunkt. Als Methoden können auch bei diesem Datentyp traditionelle oder datenverändernde Geheimhaltungsverfahren verwendet werden. Des Weiteren werden die Enthüllungsrisiken der geografischen Differenzbildung und das Prüfen bei der Darstellung von Verhältniszahlen in Karten erläutert.

1. Warum statistische Geheimhaltung?

Die Wahrung des Statistikgeheimnisses ist eine zentrale Aufgabe der amtlichen Statistik. Aus den Veröffentlichungen der statistischen Ämter darf kein Rückschluss auf die Einzelangaben der Befragten gezogen werden. Schon in der Begründung zum Bundesstatistikgesetz (BStatG)[1] 1987 wird betont, dass die statistische Geheimhaltung nicht nur eine gesetzliche Aufgabe, sondern auch essentiell für das Vertrauensverhältnis zwischen Befragten und den statistischen Ämtern ist. Dieses ist wiederum wichtig, um qualitativ hochwertige Statistiken erheben zu können (Statistische Ämter des Bundes und der Länder 2006, S. 9).

Um das Statistikgeheimnis zu wahren, werden alle Auswertungen vor der Veröffentlichung einer Prüfung unterzogen, und zwar der sogenannten statistischen Geheimhaltung. Zur Identifikation von Tabellenzellen, die möglicherweise nicht veröffentlicht werden dürfen, werden in der amtlichen Statistik üblicherweise drei Geheimhaltungsregeln angewandt. Diese Regeln können grundsätzlich auch auf Auswertungen von georeferenzierten Daten angewandt werden. Jede geografische Einheit, für die eine Information veröffentlicht werden soll, ob nach Koordinaten, nach Rasterzellen oder anderen Regionalschlüsseln, wird dabei wie ein Tabellenfeld behandelt.

In Verbindung mit dem E-Government-Gesetz (EGovG)[2] ist die Speicherung von Daten in Rastern mit einer Größe von 100 x 100 Metern erlaubt, so ist es bei der Auswertung und Veröffentlichung von Ergebnissen jedoch stets notwendig, die Geheimhaltung zu prüfen und falls notwendig durch geeignete Geheimhaltungsverfahren das Statistikgeheimnis zu wahren.

Bei Veröffentlichungen aus Bundesstatistiken gilt grundsätzlich das BstatG. Bei anderen Datenproduzenten unterliegen Auswertungen dem Bundesdatenschutzgesetz (BDSG)[3]. Hier gilt ein Geheimhaltungserfordernis nur dann, wenn die Daten personenbezogen sind. Bei Auswertungen georeferenzierter Daten ist bei diesen der Personenbezug zuerst zu prüfen. Einige Geofachdaten beziehen sich auf Sachen, wie beispielsweise Grundstücke. Wenn der jeweilige Gegenstand, der Ort oder das Grundstück einer natürlichen Person zugeordnet werden kann, sind die Veröffentlichungen nach dem BDSG zu prüfen. Unter welchen Bedingungen eine Bestimmbarkeit angenommen werden muss, ist laut Karg (2008, S. 8) letztlich nicht endgültig geklärt und in der Praxis der Aufsichtsbehörden, der Literatur und Rechtsprechung umstritten. Da bei Auswertungen von Geofachdaten aus Bundesstatistiken nach § 16 BStatG alle Einzelangaben geschützt werden müssen, ist es nicht von Bedeutung, ob die Einzelan-

1 Gesetz über die Statistik für Bundeszwecke (Bundesstatistikgesetz – BStatG) vom 22. Januar 1987 (BGBL. I S. 462, 565), zuletzt geändert durch Artikel 13 des Gesetzes vom 25. Juli 2013 (BGBL. I S. 2749).

2 Durch das Gesetz zur Förderung der elektronischen Verwaltung (E-Government-Gesetz – EGovG) vom 25. Juli 2013 (BGBL. I S. 2749) wird in § 10 BStatG (Erhebungs- und Hilfsmerkmale) die „geografische Gitterzelle" eingefügt. Diese umfasst eine Fläche von 100 x 100 Metern (= 1 Hektar oder 1 ha). Damit wurde eine Rechtsgrundlage für die Speicherung und Verbreitung georeferenzierter statistischer Angaben geschaffen. Ziel ist die flexiblere räumliche Auswertung von Statistiken. Die statistische Geheimhaltung ist aber weiterhin zu wahren. In § 13 BStatG werden auch in Adressdateien die „Geokoordinaten" ergänzt. Hierdurch erfolgte die Klarstellung, dass die dauerhafte Speicherung der Geokoordinate in Adressdateien zulässig ist.

3 Bundesdatenschutzgesetz (BDSG) in der Fassung der Bekanntmachung vom 14. Januar 2003 (BGBL. I S. 66), zuletzt geändert durch Artikel 1 des Gesetzes vom 14. August 2009 (BGBL. I S. 2814).

gaben von natürlichen oder juristischen Personen gemacht wurden oder aus Verwaltungsdaten entnommen sind.

2. Geheimhaltungsprüfung

a) Anwendung der Geheimhaltungsregeln

Die gängigen Geheimhaltungsregeln in der amtlichen Statistik sind die Mindestfallzahl-, die Randwert- und die Dominanzregel. Die Mindestfallzahlregel schreibt vor, dass stets eine festgelegte Anzahl an Befragten zu einem veröffentlichten Ergebnis beiträgt. Nach der Randwertregel ist eine Tabellenzeile oder -spalte zu sperren, wenn alle Merkmalsträger die gleiche Ausprägung ausweisen. Ein Dominanzfall liegt vor, wenn ein oder zwei Beitragende einen zu großen Anteil der Gesamtsumme auf sich vereinen.

Überträgt man die Mindestfallzahlregel auf Rasterveröffentlichungen, so sollen in jeder Gitterzelle mehrere oder kein Merkmalsträger enthalten sein. Ist diese Bedingung nicht erfüllt, darf zu dieser Gitterzelle keine Angabe veröffentlicht werden. Der Thematik, welcher Schwellenwert bei der Mindestfallzahlregel bei der Auswertung von georeferenzierten Daten gewählt werden sollte, wird in Abschnitt 2b) nachgegangen.

Die Dominanzregel ist bei den Statistiken relevant, bei denen metrische Merkmale georeferenziert vorliegen. Das trifft beispielsweise oft bei den Wirtschaftsstatistiken zu. Die Konzentration von Kennzahlen wie Umsätze, Exporte oder Anzahl der Beschäftigten bei kleinen regionalen Gliederungen kann eine leichte Zuordnung zu den beitragenden Einheiten erlauben. Um zu überprüfen, ob ein Dominanzfall vorliegt, wird die p%-Regel empfohlen. Ein Tabellenfeld wird nach der p%-Regel dann gesperrt, wenn der zweitgrößte Beitragende den größten Beitragenden aufgrund der veröffentlichten Gesamtsumme und der Kenntnis seines eigenen Beitrags so genau schätzen könnte, dass die Schätzung um weniger als p% vom tatsächlichen Wert abweicht.

Auch bei der Auswertung von georeferenzierten Daten können Randwerte entstehen. Wenn alle Merkmalsträger in einer regionalen Einheit die gleiche Ausprägung aufweisen, so ist enthüllt, welche Ausprägung jeder einzelne Merkmalsträger bei diesem Merkmal aufweist. Als Extrembeispiel stelle man sich eine Rasterkarte vor, die angibt, wie viele Unternehmen wegen Umweltvergehen verurteilt wurden. In einer Rasterzelle wurden 23 Unternehmen verurteilt, es gibt in diesem Gebiet jedoch nur 23 Unternehmen, was aus einer anderen Quelle ermittelt werden kann. Dann weisen alle Unternehmen die gleiche Ausprägung auf: „Verurteilt: ja" und die Leserinnen und Leser wissen, dass jedes einzelne Unternehmen ein Umweltvergehen begangen hat. Bei Randwert-Konstellationen bestehen Enthüllungsrisiken, auch wenn es sich nicht um zu kleine Fallzahlen handelt (Walla 2007).

Daten der amtlichen Statistik unterliegen dem BStatG. Jedoch ist eine Geheimhaltungsmaßnahme nach §16 Satz 1 Nr. 4 BStatG nur notwendig, wenn die Einzeldaten dem Merkmalsträger zugeordnet werden können. Ein Gutachten von Karg (Karg 2008), das sich auf Daten bezieht, die dem BDSG unterliegen, kann hier eventuell zu Vergleichszwecken herangezogen werden. Demnach besteht ein Personenbezug bei Fachdaten zur Landnutzung (Karg 2008, S. 61) und Kriminalität (Karg 2008, S. 65).

b) Welche Mindestfallzahl soll angewendet werden?

In einem von der Arbeitsgruppe Geodatenschutz des Interministeriellen Ausschusses für Geoinformationswesen (2014, S. 13) erarbeiteten „Behördenleitfaden zum Datenschutz bei Geodaten und -diensten" wird empfohlen, dass bei Auswertungen mit geografischem Bezug stets mindestens vier Einheiten zu einem Datum beitragen sollen. Es wird hier vor allem auf Haushalte abgestellt, aber erwähnt, dass dieses Kriterium auch bei Daten auf Personenebene angewandt werden kann. Dann sei kein Personenbezug mehr herstellbar. Allerdings enthält dieser Leitfaden außer dieser Handlungsempfehlung keine methodische Begründung dieses Schwellenwertes. Als Quelle wird vielmehr folgender Abschnitt aus Karg und Weichert (2007, S. 25) genannt:

„Eine Anonymisierung im Hinblick auf das Eigentum kann dadurch erreicht werden, dass mehrere Grundstücke zusammengefasst werden und deren Eigenschaft in gemeinsamer Form, z. B. durch einen Mittelwert, beschrieben wird. Diese Form des Zusammenfassens von personenbezogenen Daten wird auch Aggregieren genannt. Es gibt keine eindeutigen rechtlichen Vorgaben, wann von einer Anonymität hinreichend gewährleistenden Aggregierung gesprochen werden kann. Unstreitig ist, dass eine Zusammenfassung von zwei Personeneinheiten zu einer Merkmalseinheit noch nicht ausreicht. Entsprechend der statistikrechtlichen Praxis kann man davon ausgehen, dass bei einer Zusammenführung von mindestens vier Personeneinheiten zu einem Datensatz der Personenbezug hinreichend verschleiert wird."

Es ist allerdings festzustellen, dass in der Praxis und der Literatur unterschiedliche Schwellen genannt werden. Karg (2008, S. 23) dokumentiert, dass manchmal die „Aggregierung" von mindestens vier, von Aufsichtsbehörden aber manchmal bis zu zehn Grundstücken gefordert werde. Zur Begründung wird auf die erhöhte Sicherheit der Betroffenen verwiesen. Je mehr Angaben zusammengenommen werden, desto geringer sei die Wahrscheinlichkeit der Reidentifizierung. In seinem Gutachten empfiehlt der Rat für Sozial- und Wirtschaftsdaten (2012, S. 49) gar, den australischen Ansatz zu übernehmen. Dort werden die Raster so groß gewählt, dass sie mindestens 30 Wohneinheiten oder 60 Personen enthalten. Auch Szibalski (2007, S. 142) dokumentiert sehr unterschiedliche Mindestfallzahlen. Bei einer Umfrage unter nationalen Statistikämtern in Europa im Jahr 2006 schwankten diese zwischen zwei Beitragenden bis zu mindestens 20 und in einem Fall gar mindestens 31 Merkmalsträgern in einer Rasterzelle. In einem aktuelleren Beitrag auf der Konferenz des „European Forum for Geography and Statistics (EFGS)" berichten Martin et al. (2013), dass

in Norwegen eine Mindestfallzahl von 0 (also keine), in Österreich 30 und in England 100 gilt. Auch sie stellen fest, dass es in Europa derzeit eine große Variation der Mindestfallzahlen gibt. Auf der Internetseite des EFGS selbst ist ein Leitfaden[4] zu finden, der Rasterzellen mit einer Größe von 1 km² empfiehlt. Falls Geheimhaltung relevant wäre, so solle mit einer Mindestfallzahl von 10 geprüft werden.

Im Gutachten von Karg (2008, S. 24) werden weitere Randbedingungen für eine „zuverlässige Aggregierung", gemeint ist hier eine ausreichende Zusammenfassung, genannt. Diese entsprechen der üblichen Praxis in den statistischen Ämtern. Die Prüfung der Geheimhaltung bei zusammenhängenden Einheiten muss auf der hierarchischen Ebene durchgeführt werden, von der aus Eigentumsbeziehungen zu den unteren Ebenen bestehen. Um beispielsweise die Einzelangaben einer Pflegeeinrichtung zu schützen, müssen die Angaben von Einrichtungen verschiedener Träger zusammengefasst werden. Sollen Angaben über Betriebe veröffentlicht werden, müssen die Angaben von so vielen Betrieben zusammengefasst werden, dass mehrere Unternehmen als Eigentümer hinter diesen Betrieben stehen. Weichert (2009, S. 351) nennt als weiteres Kriterium, dass keine Zusatzinformationen vorhanden sein dürfen, dass beispielsweise alle Grundstücke gleich groß sind, wie es in manchen Neubausiedlungen der Fall ist. Dies würde in der amtlichen Statistik als Randwert identifiziert und sollte daher nicht veröffentlicht werden.

Eine alternative Handlungsempfehlung, die in diesem Beitrag nicht unterstützt werden kann, bezieht sich bei Auswertungen mit geografischem Bezug auf den Maßstab bzw. Detaillierungsgrad von Karten. In einem Beitrag wurde als „allgemeine Regel" ein Maßstab von 1:10 000 empfohlen (Karg 2008, S. 13). Allerdings merkt Weichert (2009, S. 350) korrekterweise an, dass es „keine maximalen Grundstücksgrößen gibt, und daher keine Flächengröße [bei Auswertungen von Flächendaten] genannt werden kann, ab der ein Personenbezug ausgeschlossen werden kann". Ein Maßstab alleine kann deshalb als Prüfkriterium bei der Sicherstellung des Statistikgeheimnisses nicht empfohlen werden.

Es wird geraten, vor der Veröffentlichung von Karten oder Rasterzellen die Besetzungszahlen jeder geografischen Einheit anhand der aufgeführten und diskutierten Geheimhaltungsregeln zu prüfen. Eine einheitliche Empfehlung lässt sich aus der Literatur nicht ableiten. Der Datenschutzbeauftragte des Landes Schleswig-Holstein Thilo Weichert (2009, S. 351) empfiehlt: „In der Praxis bietet es sich an, eine gröbere Aggregierung vorzunehmen. Bei einer Zusammenfassung von 10 Grundstücken kann in der Regel von einer hinreichend Anonymität herstellenden Aggregierung ausgegangen werden." Je höher die Mindestfallzahl, desto eher ist die Besetzung für den Datenschutz ausreichend. Allerdings ist der Informationsverlust auch höher, da durch die Anwendung der Geheimhaltungsregel mehr Gitterzellen oder Kartenflächen als schützenswert markiert werden.

c) Traditionelle Geheimhaltungsmethoden: Sperrung oder Vergröberung

Zeigt eine der Geheimhaltungsregeln an, dass die Informationen für die regionale Einheit sensibel oder kritisch ist, so darf für diese regionale Einheit die Angabe nicht veröffentlicht werden. Sind in einer Rasterzelle laut Mindestfallzahlregel zu wenige Befragte enthalten oder dominiert ein Beitragender nach der Dominanzregel, kann als Maßnahme die Information in der Zelle unterdrückt werden. Liegt ein Randwertproblem vor, müssen mehrere Felder unterdrückt werden. Das Geheimhaltungsverfahren wird Zellsperrung genannt und den traditionellen Geheimhaltungsmethoden zugeordnet.

In Tabellen werden gesperrte Felder durch den Punkt „." gekennzeichnet. In Karten wird in der Regel eine Grau- oder Weiß-Färbung verwendet und diese in der Legende erklärt. Beim Atlas „Agrarstatistik" wird sie in der Legende mit „Kein Wert vorhanden oder geheim zu halten" erläutert. Wird als Geheimhaltungsmethode die Sperrung verwendet, bleibt zu untersuchen, ob aufgrund von Summenbeziehungen zu veröffentlichten Randsummen weitere Gitterzellen zu sperren sind. Dieses Problem ergibt sich, wenn für größere Gebiete, die sich additiv aus den einzelnen Gitterzellen ergeben, ebenfalls Werte veröffentlicht werden. In diesem Fall ergibt sich bei Unterdrückung eines Wertes noch kein Schutz.

Alternativ kann das Raster lokal größer gezogen oder mehrere Zellen zusammengelegt werden. Es kann auch für die gesamte Karte ein breiteres Raster gewählt werden.

Jedes Geheimhaltungsverfahren führt zu einem Informationsverlust. Das Sperren von Werten in einzelnen Rasterzellen oder georeferenzierten Flächen in Kartendarstellungen bedeutet ein Zurückhalten von Informationen. Das Vergrößern der Flächen oder Zellen bringt einen Verlust der geografischen Genauigkeit (loss in spatial accuracy) mit sich. Eine Analyse zum Informationsverlust, der durch größere Gitterzellen entsteht, findet sich im Beitrag von Schmon in diesem Heft.

d) Datenverändernde Geheimhaltungsverfahren

Eine Alternative zu den traditionellen Geheimhaltungsverfahren, die Informationen in Veröffentlichungen unterdrücken, indem beispielweise Tabellenfelder oder Gitterraster gesperrt werden, ist die neuere Klasse der datenverändernden Geheimhaltungsverfahren. Bei diesen wird ein Rückschluss auf einzelne Befragte verhindert, indem entweder die Einzeldaten oder die statistischen Ergebnisse leicht verändert werden. Ein Überblick über klassische

4 Der Leitfaden (EFGS Standard for official grid statistics, all variables v.1.0, Wordformat) kann von folgender Seite abgerufen werden (Stand 29. 04. 2014): http://www.efgs.info/geostat/1B/efgs-standard-for-official-statistics/view. Dort findet sich folgende Textpassage: "The following statistical units, variables and divisions are recommended as official statistics; free of charge and without disclosure to be reported on km² grid cells. If disclosure is still considered to be needed, the recommended threshold value is set to 10." Die im Leitfaden aufgeführten statistischen Auswertungen fallen in Deutschland alle unter § 16 BStatG, sobald Bundesstatistiken ausgewertet werden, um die Kennzahlen zu berechnen.

und datenverändernde Geheimhaltungsverfahren findet sich in Höhne (2010).

In manchen Ländern werden datenverändernde Geheimhaltungsverfahren zur Sicherung des Statistikgeheimnisses bei Veröffentlichungen, die aus georeferenzierten Daten erstellt werden, diskutiert oder verwendet. Einen Überblick über Vorschläge, datenverändernde Verfahren auf Geodaten anzuwenden, bietet Young et al. (2012). Armstrong et al. (1999) hat dabei einen neuen Begriff für die Anwendung der Datenveränderung auf die georeferenzierten Merkmale eingeführt: „geographical masking" oder verkürzt zu „geomasking" bezeichnet die prätabulare, datenverändernde Geheimhaltung durch Veränderung der geografischen Koordinaten.

In der Praxis wurde in Finnland ein selbst weiterentwickeltes Verfahren zur „restringierten lokalen Imputation" eingesetzt (Statistics Finland 2010). In England wurde eine Variation des Record Swapping vorgeschlagen; bei diesem sogenannten „local density swapping" wird in dichter besiedelten Regionen mit einer geringeren Wahrscheinlichkeit „geswappt", also das Attribut der geografischen Verortung mit anderen Merkmalsträgern getauscht. Das Verfahren wurde an anonymisierten Zensusdaten von 1991 getestet und schnitt bei den Informationsverlustmaßen besser ab als ein zufälliges Record Swapping Verfahren. In der Gesamtpopulation einzigartige Ausprägungen können durch eine zufällige Veränderung der geografischen Verortung durch ein Swapping-Verfahren jedoch nicht ausreichend geschützt werden (Young et al. 2012).

In Deutschland wurde ein Mikroaggregationsverfahren, das sich auch als prätabulares Geheimhaltungsverfahren eignen würde, auf georeferenzierte Arbeitsmarktdaten angewandt. In erster Linie wurden die Gruppen von mindestens 15 Erwerbstätigen, jedoch nicht aus Geheimhaltungsgründen, gebildet, sondern um sogenannte Nachbarschaften zu bilden und dort Nachbarschaftseffekte zu untersuchen (Scholz et al. 2012).

Werden georeferenzierte Ergebnisse als detaillierter Input für weitere Modellrechnungen und nicht direkt zur Veröffentlichung benötigt, so besteht eine Möglichkeit darin, der Wissenschaft in geschützten Räumen und unter organisatorisch-rechtlichen Rahmenbedingungen über Forschungsdatenzentren Zugang zu „georeferenzierten Mikrodaten" bzw. zu georeferenzierten Ergebnissen ohne Geheimhaltungsprüfung zu ermöglichen. Dann kann die Wissenschaft innerhalb des Forschungsdatenzentrums forschen, und die dort erzeugten Ergebnisse werden danach auf das Statistikgeheimnis geprüft.

3. Aspekte des Rasters aufgrund der Geheimhaltung

a) Gitterzellenbasierte Grundeinheiten

Um den gespeicherten Datensatz flexibel auswerten zu können, sollten die Gitterlänge, die in Veröffentlichungen verwendet wird, und die Basislänge des Gitterrasters, das an den Einzeldaten gespeichert ist, in einer Beziehung zueinander stehen. Aufgrund der Änderungen des §10 BStatG darf an den Einzeldaten von Bundesstatistiken die Zugehörigkeit zu einer Gitterzelle mit der Größe 100 x 100 Meter gespeichert werden. Wenn sich bei Auswertungen herausstellt, dass dieser Detailgrad häufig nicht für Veröffentlichungen verwendet werden kann, so sollte eine „übliche" Gitterlänge für diese Statistik festgelegt werden.

Durch das im Rahmen von INSPIRE festgelegte Euro-Grid, bei dem sowohl die Projektion (Lambert Azimuthal Equal Area), die Lage der Gitterzellen, die Gitterweite und Schlüssel als Identifikatoren jeder Zelle etabliert wurden, sind neue funktionale Raumbezüge geschaffen worden. Die eindeutigen Schlüssel können als RasterIDs verwendet werden. Dies ist hilfreich, um eine feste Gebietssystematik an gitterzellenbasierten Grundeinheiten und somit Regionsidentifikatoren zu dokumentieren. Dadurch können Raumstrukturdaten (z. B. Bebauung, Erreichbarkeit) in Forschungsvorhaben weiter verwendet werden. Auch können beispielsweise Wissenschaftler Geofachdaten über diese RasterIDs an Mikrodaten in den Forschungsdatenzentren anspielen.

Solch eine rasterbasierte Gebietssystematik sollte einerseits möglichst kleinräumig sein. Andererseits sollten die Gitterzellen jedoch so ausgewählt sein, dass aufgrund des Datenschutzes für möglichst viele Gitterzellen auch Informationen enthalten sein können (Sigismund 2014). Die Vorteile rasterbasierter Gebietssystematiken sind, dass sie zeitreihenrobust und themenneutral sind. Da sie weder topografische, noch siedlungsstrukturelle / lebensweltliche Faktoren oder verwaltungstechnische Gliederungen berücksichtigen, eignen sie sich sowohl für Sozial-, Wirtschafts- als auch Bevölkerungsdaten. Die gerade genannten themenbasierten Gliederungen eignen sich zwar besser für spezielle Planungs- und Analysezwecke im Themengebiet. Diese Gliederungen sind aber meistens zeitlichen Änderungen unterworfen. Rasterbasierte Gebietssystematiken sind dagegen zeitlich stabil. Damit sind sowohl zeitreihenorientierte aber auch themenübergreifende Analysen problemlos möglich. Bei anderen Gebietssystematiken bedingen die Pflege zeitlicher Veränderungen bei gleichen Systematiken bzw. die Abstimmung verschiedener Systematiken einen erheblichen Vorbereitungsaufwand bei der Analyse der geobasierten Daten.

b) Informationsverlust minimieren durch hierarchische Gitterstruktur

Grundsätzlich ist ein bundesweit einheitliches Raster wünschenswert. In allen Gebieten die gleiche Rastergröße zu verwenden, hat jedoch auch eine Reihe von Nachteilen. Bei kleinmaschiger Auswertung entstehen in dünn besiedelten Landgebieten oft viele Geheimhaltungsfälle; wählt man ein großes Raster, ist der Informationsgehalt für Städte sehr niedrig. Das Informationspotenzial der Daten wird dann nicht genutzt.

Fasst man benachbarte Gitterzellen bis zur Erreichung von Mindestbesetzungszahlen zu Nachbarschaftsarealen zusammen, entstehen unregelmäßige „Gitterklumpen". Das Verfahren ist methodisch aufwändig und nicht zeitreihenrobust (Sigismund 2014).

Optimal wäre daher eine Gebietssystematik, deren Gitterbreite variiert. Hier würde sich eine hierarchische Aggregation anbieten. Eine solche wird vom Bundesamt für Statistik der Schweiz (Meyer 2011) und in Österreich (Strobl 2005) bereits verwendet. Anhand eines Datenbestands der Gebäude in Deutschland haben Behnisch et al. (2013) mit diesem Verfahren Mischrasterkarten für Hamburg erstellt. Bei der hierarchischen Aggregation wird von einer kleinen Basislänge der Gitterzellen gestartet. Sind die Besetzungszahlen nicht ausreichend, werden hierarchisch vier Gitterzellen zu einer neuen Zelle mit doppelter Gitterlänge zusammengefasst. Sollten auch in diesen größeren Zellen noch Mindestbesetzungszahlen unterschritten sein, so werden wiederum vier Zellen zu einer sehr großen zusammengelegt, die dann eine Gitterlänge vom Vierfachen der Basislänge hat.

Bei der Interpretation ist zu beachten, dass die Flächen der Gitterzellen nun unterschiedlich groß sind und die Absolutwerte daher nicht mehr zugleich auch Dichtewerte darstellen.

4. Geografische Differenzbildung

Die Mehrheit der Auswertungen der amtlichen Statistik bezieht sich auf administrative Gebiete (wie z. B. Länder, Regierungsbezirke, Kreise oder Gemeinden). Aber in einigen Statistiken gibt es bereits abweichende regionale Gliederungssysteme. Die Reisegebiete in der Tourismusstatistik, die Wassereinzugsgebiete bei Wasserstatistiken, wie der öffentlichen Abwasserbehandlung, oder auch die EU-Orte (Siedlungsstruktur[5]) beim Zensus 2011 (Heidrich-Riske et al. 2013, S. 474) weichen von den administrativen Einheiten ab[6]. Bei diesen Statistiken wird eine andere, funktional besser passende räumliche Abgrenzung gewählt als die administrativen Einheiten. Wenn Statistiken wie die Wasserstatistiken oder die Tourismusstatistik zusätzlich zur funktionsräumlichen Gliederung auch für administrative Gebiete ausgewertet werden, können durch das „Übereinanderlegen" verschiedener Gliederungssystematiken – das sogenannte Verschneiden – kleine Schnittmengen entstehen. Die Thematik wird auch als „disclosure by geographical differencing" in der internationalen Geheimhaltungsliteratur diskutiert (Hundepool et al. 2010, S. 169).

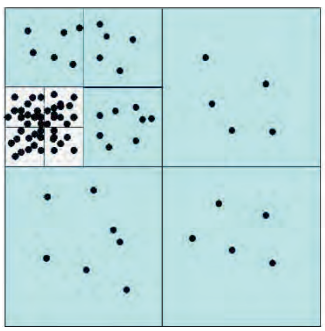
Dieses Enthüllungsrisiko ist zusätzlich zu ausreichenden Besetzungszahlen zu prüfen. Durch das Verschneiden zweier Gebietsklassifikationen entsteht eine neue, deutlich detailliertere Gliederung. Diese können sich Leserinnen und Leser, denen beide Publikationen zur Verfügung stehen, selbst berechnen. In der Geheimhaltung wird in diesem Falle auch von Restkategorien gesprochen. Die Besetzungszahlen und die Verteilung von Einheiten müssen auch für diese neu entstehende detailliertere Gliederungssystematik ausreichend sein. Sonst dürfen aufgrund der Datenschutzerfordernisse die Ergebnisse nicht nach beiden Systematiken gleichzeitig veröffentlicht werden.

Das Risikoszenario ist in Abbildung 3 dargestellt. Werden gleichzeitig Ergebnisse für die Gebietseinheit der besiedelten Gemeindeflächen und für eine andere regionale Gliederung, hier Rasterzellen, veröffentlicht, so können bei Überlappungen dieser beiden Gliederungen durch die Differenzbildung Enthüllungsrisiken entstehen. Indem man vom

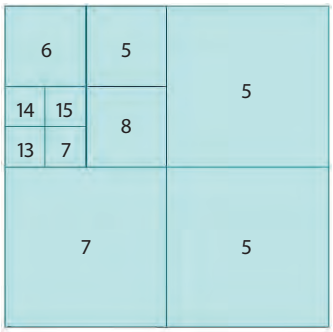
5 „Auf administrativen Grenzen basierende Zahlen zur Bevölkerungsdichte beziehen immer die gesamte Fläche einer regionalen Einheit mit ein, wodurch das Ergebnis von verfügbarem Raum je Einwohner unter Umständen von der Realität vor Ort abweichen kann. Das Konzept der ‚EU-Orte' vermag somit die Realität ein Stück weit näher abzubilden, indem es die tatsächlich gewachsenen Siedlungsflächen in den Fokus rückt." (Heidrich-Riske et al. 2013, S. 474f.)

6 Die Aufzählung erhebt keinen Anspruch auf Vollständigkeit.

1 | Prinzip der hierarchischen Aggregation



2 | Ergebnis der hierarchischen Aggregation



Quelle: Sigismund 2014

Wert der Gemeinde A den Wert der Gitterzelle 2 subtrahiert, so erhält man den Wert des Anteils der Gemeinde A in der Gitterzelle 1. In diesem Szenario würde bei zu geringer Besetzungszahl, einem Randwert oder einem dominierenden Beitragenden die Sensitivität dieses Gemeindeteils bereits auffallen, wenn die Geheimhaltungsprüfung für Rasterzelle 1 durchgeführt wird. Der Wert für Gitterzelle 1 würde bei der Verletzung einer der Geheimhaltungsregeln nicht veröffentlicht.

Ein leicht verändertes Szenario ist in Abbildung 4 dargestellt. In diesem Fall kann das Ergebnis der unabhängigen Geheimhaltungsprüfung der Gitterzellen und der Gemeinden sein, dass keine Geheimhaltungsfälle auftreten. Wurden jedoch bereits Gemeindeergebnisse veröffentlicht, so kann durch die Subtraktion des Wertes der Gemeinde B vom Wert der Gitterzelle 1 auch hier der Wert für den Teil der Gemeinde A berechnet werden, der in Gitterzelle 1 liegt. Der Wert dieses Teils der Gemeinde A entsteht als Schnittmenge der beiden regionalen Gliederungen. Für diese Schnittmenge müssten wieder ebenfalls alle Geheimhaltungsregeln geprüft werden. Ergebnisse nach beiden regionalen Gliederungen dürfen zusammen nur veröffentlicht werden, wenn in allen möglichen Schnittmengen keine Geheimhaltungsfälle auftreten.

5. Darstellungen von Verhältniszahlen in Karten

Sollen Ergebnisse von Geofachdatenauswertungen in Karten dargestellt werden, so wird oft die Einfärbung in unterschiedlichen Farbtönen gewählt. Dies entspricht dem Nachweis eines Ergebnisses in verschiedenen Kategorien, beispielsweise niedrig, mittel, hoch. Aufgrund einer begrenzten Anzahl an Farbtönen oder -schattierungen werden keine genauen Werte, sondern Intervalle, von denen in der Regel die Grenzen veröffentlicht werden, publiziert. Dies gilt natürlich nicht bei interaktiven Karten, bei denen der exakte Wert für eine geografische Einheit in der Regel angezeigt wird, z. B. als Mouseover-Effekt, sobald man mit dem Mauszeiger über die Einheit fährt oder durch das zusätzliche Anzeigen von Grafiken oder Tabellen, wenn man diese anklickt.

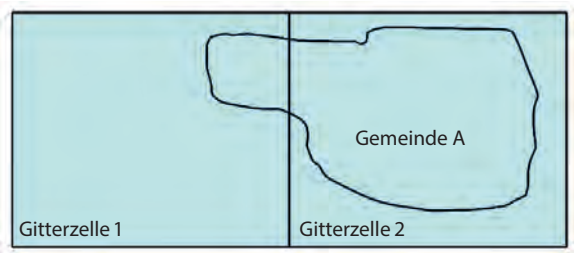
Grundlage der Kartendarstellung sind meist Verhältniszahlen. Das liegt darin begründet, dass viele Merkmale stark korreliert sind, sodass eine Darstellung von Absolutwerten verschiedener Themen kaum einen Informationsgewinn ermöglicht. Wird beispielsweise die Anzahl an Einwohnern kartiert, so ergibt sich ebenfalls eine gleiche räumliche Struktur bei der Kartierung von Seniorinnen und Senioren, Kindern usw. Es wird letztendlich nur die Information „dünn oder dicht besiedeltes Gebiet", „Dorf oder Stadt" abgebildet. Die Kartierung des Anteils der Seniorinnen und Senioren an der Einwohneranzahl (Altenquotient) oder andere Anteilsinformationen sind dagegen viel aussagekräftigere Informationen. Verhältniszahlen bergen dabei als Quotient aus zwei Einzelwerten das Risiko, dass mit der Kenntnis eines Einzelwertes und einer hinreichenden Genauigkeit der Verhältniszahl auch der andere Einzelwert rückgeschlossen werden kann. Handelt es sich bei einem der Werte um eine geheim zu haltende Information, so besteht deshalb ein Geheimhaltungsrisiko. Diesem kann nur durch eine Vergröberung oder Geheimhaltung der Verhältniszahl begegnet werden. Das Geheimhaltungsproblem für Verhältniszahlen kann auch bereits auftreten, wenn man nur sehr grobe Informationen über den einen Einzelwert hat.

Beispiel:

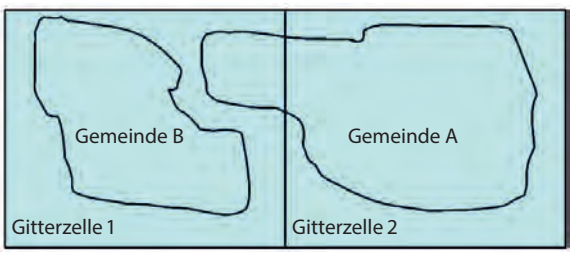
Region	Ärzte je 1000 Einwohner	Einwohner
Gemeinde C.....	0,8	ca. 1 000 bis 1 500

Wird für eine Gemeinde oder eine Rasterzelle als Kennzahl ausgewiesen, dass dort 0,8 Ärzte je 1000 Einwohner praktizieren, so kann man trotz nur sehr grober Kenntnis der Einwohnerzahl berechnen, wie viele Ärzte absolut in der Region ermittelt wurden. Obwohl mit der Veröffentlichung von Verhältniswerten keine direkten Einzelangaben veröffentlicht werden, enthalten sie für kleine Einheiten ein großes Potenzial, einzelne Angaben zurückzuschließen. Das liegt darin begründet, dass neben der notwendigen Kenntnis von einem der beiden Werte bei vielen Merkmalen die Ganzzahligkeit der Einheiten unterstellt werden kann (z. B. Personen, Unternehmen usw.). Damit ergeben sich relativ große Intervalle, in denen noch ein Rückschluss auf die ganzzahligen Werte möglich ist. Selbst bei nicht ganzzahligen Einzelangaben ergeben sich noch hohe Risiken.

3 | Risikoszenario 1 der regionalen Differenzbildung



4 | Risikoszenario 2 der regionalen Differenzbildung



Quelle: eigene Darstellung

6. Zusammenfassung

Die Wahrung des Statistikgeheimnisses ist für die amtliche Statistik eine zentrale Aufgabe. Um den Schutz der für eine Bundesstatistik gemachten Einzelangaben zu wahren, werden die Geheimhaltungsregeln geprüft und bei Geheimhaltungsfällen ein Geheimhaltungsverfahren angewendet. Jedes Verfahren führt dabei immer zu einem Informationsverlust in den geplanten Veröffentlichungen, schützt aber dafür die Einzelangaben. Der vorliegende Beitrag stellte die wichtigsten Aspekte bei der Geheimhaltungsprüfung von Veröffentlichungen aus georeferenzierten Daten zusammen. Es wurde die Anwendung der Geheimhaltungsregeln auf geplante Publikationen von Geofachdaten beschrieben. Dabei wurde deutlich, dass es für die Mindestfallzahlregel in der Anwendung noch keine einheitliche Mindestfallzahl gibt. Neben den klassischen Geheimhaltungsverfahren der Sperrung und Vergröberung werden in der Literatur auch datenverändernde Geheimhaltungsverfahren bei Geodaten vorgeschlagen, die in manchen Ländern bereits eingesetzt wurden. Die Besonderheiten von Verhältniszahlen in Karten und das Problem der geografischen Differenzbildung, das bei zusätzlichen Rasterkarten in noch stärkerem Maße auftritt, wurden erläutert. Dennoch wird deutlich, dass die statistische Geheimhaltung bei georeferenzierten Daten noch nicht abschließend untersucht und geklärt ist.

Weichert (2009, S. 352) stellt dabei die Forderung auf, dass „die Geodaten verarbeitenden Stellen in Verwaltung und Wirtschaft […] durch einen sensiblen Umgang dazu beitragen [können], dass das Vertrauen in die Wahrung des Persönlichkeitsrechts nicht beschädigt wird." Im Bereich der wissenschaftlichen Diskussion des Datenschutzrechts wird des Öfteren nach Selbstverpflichtungen gerufen. Dieser bedarf es in Teilgebieten, in denen es keine rechtlichen Regelungen gibt. Im Bereich der Bundesstatistik ist diese Forderung nicht passend, da das Bundesstatistikgesetz (BStatG) für georeferenzierte Daten gleichermaßen wie bei allen anderen Daten auch gilt. Aber einheitliche und abgestimmte Handlungsempfehlungen auf Basis von methodischen Untersuchungen und rechtlichen Stellungnahmen wären für die praktische Arbeit hilfreich.

Da georeferenzierte Daten aber nicht nur in immer größeren Maße bei der amtlichen Statistik verfügbar sein werden, sondern auch von anderen Datenproduzenten erzeugt und genutzt werden, sollte die Diskussion um den „richtigen" Datenschutz bei Auswertungen von georeferenzierten Daten auch im breiteren gesellschaftlichen Rahmen geführt werden. Gutman und Stern (2007) fordern, dass Universitäten und Fachgesellschaften sich engagieren sollten, dass Wissenschaftlerinnen und Wissenschaftler im verantwortungsvollen Umgang mit georeferenzierten Daten geschult werden. Es sollen Normen und Handlungsanweisungen zur ethisch korrekten Verwendung solcher Daten entwickelt werden.

Dr. Jörg Höhne leitet die Abteilung *Gesamtwirtschaft* im Amt für Statistik Berlin-Brandenburg. Er studierte Statistik und Wirtschaftsmathematik in Berlin und Moskau und promovierte 2009 an der Universität Tübingen mit einer Arbeit über „Verfahren zur Anonymisierung von Einzeldaten".

Julia Höninger, Diplom-Volkswirtin, leitet das Referat *Volkswirtschaftliche Gesamtrechnungen, Erwerbstätigkeit*. Zuvor arbeitete sie als wissenschaftliche Mitarbeiterin im Referat *Mikrodaten, Analysen, Forschungsdatenzentrum* des Amtes für Statistik Berlin-Brandenburg.

7. Literatur

- Behnisch, Martin; Meinel, Gotthard; Tramsen, Sebastian; Diesselmann, Markus (2013): Using quadtree representations in building stock visualization and analysis. Erdkunde Vol. 67 No. 2, S. 151-166.
- Dorer, Peter; Mainusch, Helmut; Tubies, Helga (1988): Bundesstatistikgesetz. Verlag C. H. Beck.
- Gutman, Myron P.; Stern, Paul C. (2007, Hrsg.): Putting people on the map – protecting confidentiality with linked social-spatial data. Issues arising from the integration of remotely sensed and self-identifying data. National Research Council of the National Academies – The National Academic Press, Washington D.C.
- Heidrich-Riske, Holger; Scholz, Bettina; Stepien, Halina (2013): GIS-gestützte Ermittlung der „EU-Orte" im Rahmen des Zensus 2011 für die Datenlieferung an Eurostat. Statistisches Bundesamt, Wirtschaft und Statistik, Juli 2013, S. 467-475.
- Höhne, Jörg (2010): Überblick über Anonymisierungsverfahren für Mikrodaten. Kapitel 2, in: Verfahren zur Anonymisierung von Einzeldaten. Statistik und Wissenschaft, Band 16, Statistisches Bundesamt, Wiesbaden, S. 22-45. Verfügbar unter: https://www.destatis.de/DE/Publikationen/StatistikWissenschaft/Band16_Anonymisierung-Einzeldaten_1030816109004.pdf?__blob=publicationFile
- Hundepool, Anco; Domingo-Ferrer, Josep; Franconi, Luisa; Giessing, Sarah; Lenz, Rainer; Naylor, Jane; Schulte Nordholt, Eric; Seri, Giovanni; de Wolf, Peter-Paul (2010): Handbook on Statistical Disclosure Control. Version 1.2, available from http://neon.vb.cbs.nl/casc/SDC_Handbook.pdf
- Interministerieller Ausschuss für Geoinformationswesen (2014): Behördenleitfaden zum Datenschutz bei Geodaten und -diensten, http://www.imagi.de/SharedDocs/Downloads/IMAGI/DE/Imagi/behoerdenleitfaden.pdf?__blob=publicationFile

- Karg, Moritz (2008): Datenschutzrechtliche Rahmenbedingungen für die Bereitstellung von Geodaten für die Wirtschaft. Gutachten im Auftrag der GIW-Kommission. Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein (ULD).
- Karg, Moritz; Weichert, Thilo (2007): Datenschutz und Geoinformationen. Eine Studie im Auftrag des Bundesministeriums für Wirtschaft und Technologie (BMWi). Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein (ULD).
- Martin, David; Steinnocher, Klaus; Petri, Ekkehard (2013): Sensitivity analysis of disclosure control measures. EFGS conference, Sofia, verfügbar unter: http://www.efgs.info/geostat/workshops/efgs-2013-sofia-bulgaria/efgs-2013-conference-1/day2_ws1_8_presentation_martin
- Meyer, Werner (2011): Arbeitsergebnisse des Bundesamts für Statistik der Schweiz. Vortrag auf dem 3. Dresdner Flächensymposium 2011.
- Rat für Sozial- und Wirtschaftsdaten (2012): Endbericht der AG „Georeferenzierung von Daten" des RatSWD – Bericht der Arbeitsgruppe und Empfehlung des Rates für Sozial- und Wirtschaftdaten (RatSWD) http://www.ratswd.de/Geodaten/downloads/RatSWD_Endbericht_Geo-AG.pdf
- Scholz, Theresa; Rauscher, Cerstin; Reiher, Jörg; Bachteler, Tobias (2012): Geocoding of German Administrative Data – The Case of the Institute for Employment Research, FDZ-Methodenreport 9/2012, verfügbar unter: http://doku.iab.de/fdz/reporte/2012/MR_09-12_EN.pdf
- Statistics Finland (2010): Production and dissemination of grid data since the 1970 Census in Finland. Conference of European Statisticians, Fifty-eighth plenary session, Paris, 8-10 June 2010.
- Statistische Ämter des Bundes und der Länder (2006): http://www.statistikportal.de/Statistik-Portal/qualistandards.pdf
- Strobl, Josef (2005): Hierarchische Aggregation: Detailinformation versus Datenschutz am Beispiel adressbezogen georeferenzierter Datensätze. Salzburger Geographische Arbeiten 38, Salzburg, S. 163-171.
- Szibalski, Martin (2005): Anonymität von Erhebungseinheiten und statistische Geheimhaltung in digitalen Karten amtlicher Statistikdaten. Methoden – Verfahren – Entwicklungen Heft 2/2005, Statistisches Bundesamt, S. 5-7.
- Szibalski, Martin (2007): Kleinräumige Bevölkerungs- und Wirtschaftsdaten in der amtlichen Statistik Europas – Ergebnisse einer Umfrage zur Speicherung, Analyse und Publikation. Wirtschaft und Statistik 2/2007, Statistisches Bundesamt, S. 137-144.
- Walla, Wolfgang (2007): Standpunkt: Was ist daran geheim? Statistisches Monatsheft Baden-Württemberg 8/2007, S. 51-53.
- Weichert, Thilo (2009): Geodaten – datenschutzrechtliche Erfahrungen, Erwartungen und Empfehlungen. Datenschutz und Datensicherheit, Heft 6/2009, S. 347-252.
- Young, Caroline; Martin, David; Skinner, Chris (2009) Geographically intelligent disclosure control for flexible aggregation of census data, International Journal of Geographical Information Science, 23:4, S. 457-482, verfügbar unter: http://dx.doi.org/10.1080/13658810801949835