Enderle, Tobias; Vollmar, Meike

Article

Geheimhaltung in der Hochschulstatistik

WISTA - Wirtschaft und Statistik

Provided in Cooperation with:
Statistisches Bundesamt (Destatis), Wiesbaden

Suggested Citation: Enderle, Tobias; Vollmar, Meike (2019) : Geheimhaltung in der Hochschulstatistik, WISTA - Wirtschaft und Statistik, ISSN 1619-2907, Statistisches Bundesamt (Destatis), Wiesbaden, Vol. 71, Iss. 6, pp. 87-98

This Version is available at:
https://hdl.handle.net/10419/213051

GEHEIMHALTUNG IN DER HOCHSCHULSTATISTIK

Dr. Tobias Enderle, Dr. Meike Vollmar

↘ **Schlüsselwörter:** Geheimhaltungsverfahren – datenveränderndes Verfahren – stochastische Überlagerung – Cell-Key-Methode – Hochschulstatistik

ZUSAMMENFASSUNG

Mit der Novelle des Hochschulstatistikgesetzes von 2016 wurde die rechtliche Grundlage für den Aufbau einer zentralen Auswertungsdatenbank geschaffen. Diese soll dazu genutzt werden, Einzelangaben der Hochschulstatistiken zu speichern und für die Hochschulplanung sowie die Bildungsberichterstattung flexibel auszuwerten. Die Sicherstellung des Statistikgeheimnisses in der Auswertungsdatenbank wird dabei automatisiert über die sogenannte Cell-Key-Methode erfolgen. Dieses datenverändernde Geheimhaltungsverfahren ermöglicht eine hohe Flexibilität bei der Auswertung und zugleich hohen Schutz vor Aufdeckung von Einzelangaben sowie geringen Informationsverlust. Der Artikel stellt das Verfahren vor, beschreibt dessen Entwicklung für die spezifischen Anforderungen der Hochschulstatistik sowie die geplante Umsetzung.

↘ ***Keywords:*** *statistical disclosure control – data perturbation method – stochastic perturbation – cell-key method – higher education statistics*

ABSTRACT

The amendments to the Higher Education Statistics Act of 2016 created the legal basis for setting up a central evaluation database. In this database, individual data of higher education statistics can be stored and flexibly evaluated for the purpose of higher education planning and educational reporting. Ensuring statistical confidentiality in the evaluation database will be automated through the so-called cell-key method. This perturbative confidentiality method allows high evaluation flexibility and, at the same time, high protection against disclosure of individual data as well as low information loss. The article presents the procedure, describes its development for the specific requirements of higher education statistics and the planned implementation.

Dr. Tobias Enderle

ist Volkswirt und hat an der Universität Trier zum Thema Imputation promoviert. Er war in verschiedenen Funktionen bei GESIS angestellt, zuletzt als kommissarischer Teamleiter in der Abteilung „Survey Design and Methodology“. Im Referat „Statistische Geheimhaltung“ des Statistischen Bundesamtes befasst er sich als Referent mit Fragen der mathematisch-statistischen Geheimhaltungsmethodik.

Dr. Meike Vollmar

ist promovierte Sozialwissenschaftlerin und seit 2010 im Statistischen Bundesamt in verschiedenen Bereichen der Bildungsstatistiken tätig. Seit Mai 2017 ist sie als Referentin im Referat „Hochschulen“ unter anderem für das Geheimhaltungsverfahren der Hochschulstatistiken zuständig.

1

Einleitung

Mit der Novellierung des Hochschulstatistikgesetzes[1] 2016 wurde das Statistische Bundesamt mit der Einrichtung und dem Betrieb einer zentralen Auswertungsdatenbank für die Hochschulstatistiken beauftragt. In dieser Datenbank sollen die Einzelangaben der Hochschulstatistiken gespeichert werden. Diese sollen so für Standard- und Sonderauswertungen im Rahmen der Hochschulplanung und -steuerung sowie für die Bildungs- und Forschungsberichterstattung zur Verfügung stehen.

Nach § 16 Bundesstatistikgesetz[2] sind Einzelangaben über persönliche und sachliche Verhältnisse, die für eine Bundesstatistik gemacht werden, von den jeweils durchführenden statistischen Stellen geheim zu halten, soweit nichts anderes bestimmt ist. Dies wird auch als Statistikgeheimnis bezeichnet.

Das Statistikgeheimnis steht in einem engen Zusammenhang mit der Auskunftspflicht nach § 15 Bundesstatistikgesetz. Die amtliche Statistik kann die erforderlichen Informationen einfordern, wenn die die Bundesstatistik anordnende Rechtsvorschrift dies festlegt. Im Gegenzug ist sie gemäß § 16 Bundesstatistikgesetz verpflichtet, die erhaltenen Informationen zu schützen, sodass sie keine Rückschlüsse mehr auf die betreffende Person und den dargelegten Sachverhalt enthalten. Damit schützt sie zugleich das Recht jedes beziehungsweise jeder Einzelnen auf informationelle Selbstbestimmung. Insofern ist die Geheimhaltung „konstitutiv" für die amtliche Statistik und eine ihrer wichtigsten Aufgaben.

Der Schutz von hochschulstatistischen Einzelangaben in Veröffentlichungen erfolgt bisher in der Regel manuell über eine Sperrung von Feldern in den einzelnen Ergebnistabellen der statistischen Ämter. Der Einsatz einer zentralen Auswertungsdatenbank erfordert allerdings ein tabellenübergreifendes, einheitliches und voll automatisiertes Geheimhaltungsverfahren.

Um ein hierfür geeignetes Verfahren zu identifizieren, wurden im Vorfeld unterschiedliche Optionen analysiert. Dieser Artikel stellt die Vor- und Nachteile der verschiedenen betrachteten Geheimhaltungsverfahren dar. Er begründet die Entscheidung für die stochastische Überlagerung nach der Cell-Key-Methode als Geheimhaltungsverfahren für die Hochschulstatistiken und erläutert die Analysen zur Bestimmung der spezifischen Parameter des Geheimhaltungsverfahrens.

2

Status quo der Geheimhaltung in den Hochschulstatistiken

In den nicht monetären Hochschulstatistiken[3] wird derzeit die Geheimhaltung der Ergebnisse in Bund und Ländern unterschiedlich umgesetzt. Das Statistische Bundesamt setzt wie einige Statistische Landesämter ein Zellsperrverfahren auf Basis des sogenannten Randsummenverfahrens[4] zur Erkennung der Geheimhaltungsfälle ein. Mittels dieses Verfahrens werden bestimmte Informationen unterdrückt, das heißt nicht ausgewiesen. Die Zellsperrung erfolgt in der Regel manuell und begrenzt auf die jeweilige Tabelle.

Generelle Anforderungen an ein Geheimhaltungsverfahren (Rohde und andere, 2018) sind eine möglichst hohe Sicherheit gegen Aufdeckung bei gleichzeitig möglichst hoher Qualität der Ergebnisse, das heißt es wird nur so viel Veränderung wie nötig vorgenommen. Denn eine Veränderung der Originaldaten erhöht zwar die Schutzwirkung, führt aber gleichzeitig zu einem Informations- und Qualitätsverlust der statistischen Ergebnisse. Der Schutz steht somit in einem Zielkonflikt mit der Ergebnisqualität, die wiederum wesentliche Voraussetzung für den Nutzen beziehungsweise die Verwendbarkeit der Daten ist.

1 Gesetz über die Statistik für das Hochschulwesen sowie für die Berufsakademien (Hochschulstatistikgesetz – HStatG) vom 2. November 1990 (BGBl. I Seite 2414), das zuletzt durch Artikel 3 des Gesetzes vom 7. Dezember 2016 (BGBl. I Seite 2826) geändert worden ist.

2 Gesetz über die Statistik für Bundeszwecke (Bundesstatistikgesetz – BStatG) in der Fassung der Bekanntmachung vom 20. Oktober 2016 (BGBl. I Seite 2394), das zuletzt durch Artikel 10 Absatz 5 des Gesetzes vom 30. Oktober 2017 (BGBl. I Seite 3618) geändert worden ist.

3 Für die monetären Hochschulstatistiken (Hochschulfinanzstatistiken) ist keine Geheimhaltung der Ergebnisse erforderlich, da das Hochschulstatistikgesetz Angaben für einzelne Hochschulen ausdrücklich erlaubt.

4 Beim Zellsperrverfahren werden einzelne Zellen von Ergebnistabellen nicht ausgewiesen (gesperrt), denen ein Re-Identifikationspotenzial zugeschrieben wird.

Die Akzeptanz des verwendeten Geheimhaltungsverfahrens bei den Nutzerinnen und Nutzern hängt nicht nur davon ab, dass die Verwendbarkeit der Daten durch das Geheimhaltungsverfahren möglichst wenig eingeschränkt wird. Entscheidend ist auch, dass Sinn und Zweck der Geheimhaltung sowie die Wirkungsweise des gewählten Verfahrens nachvollziehbar sind. Zudem können die Nutzerinnen und Nutzer die Ergebnisse nur dann korrekt interpretieren, wenn das Verfahren verständlich dokumentiert, begründet und erläutert ist.

Mit der Auswertungsdatenbank zur Hochschulstatistik lassen sich die Einzelangaben flexibel kombinieren und auswerten; ihre Implementierung stellt daher neue Anforderungen an die Geheimhaltung. Erforderlich ist hierfür ein in Bund und Ländern abgestimmtes, einheitliches und vollständig automatisierbares Geheimhaltungsverfahren, das auch tabellenübergreifend zu konsistenten Ergebnissen führt und dabei ausreichenden Schutz vor der Aufdeckung von Einzelangaben gewährleistet.

Werden in einer dezentralen Statistik wie der Hochschulstatistik die Ergebnisse nicht nur dezentral aufbereitet, sondern auch veröffentlicht, muss das Geheimhaltungsverfahren zudem in Bund und Ländern einheitlich angewendet werden (können). Ansonsten besteht die Gefahr, dass sich die Geheimhaltungsmaßnahmen gegenseitig aufheben und etwa durch Differenzenbildung zwischen Bundes- und Ländertabellen geheim gehaltene Ergebnisse aufgedeckt werden können. Nur eine einheitliche Anwendung garantiert zudem die Konsistenz und Reproduzierbarkeit der Ergebnisse und somit das Vertrauen der Nutzerinnen und Nutzer in die Hochschulstatistik.

Eine zusätzliche, spezifische Anforderung der Hochschulstatistik an das Geheimhaltungsverfahren besteht darin, dass die Geheimhaltung der Länderergebnisse und der Bundesergebnisse (zeitlich) unabhängig voneinander erfolgen kann. Ansonsten müssten die Länder mit der Veröffentlichung ihrer Ergebnisse warten, bis die Geheimhaltung auf Bundesebene abgeschlossen ist.

Auch die Vorgabe des Hochschulstatistikgesetzes, die statistischen Ergebnisse in einer Auswertungsdatenbank für flexible Auswertungen bereitzuhalten, ist eine Anforderung an das Geheimhaltungsverfahren. Dieses muss in der Datenbank integriert sein und vollständig automatisiert ablaufen. Gleichzeitig sollte der Implementierungsaufwand angesichts der zeitlichen Erwartungen der Nutzerinnen und Nutzer an den Produktivstart der Auswertungsdatenbank möglichst gering sein.

3

Auswahl des Geheimhaltungsverfahrens

Um das Geheimhaltungsverfahren zu identifizieren, das die vorgenannten Anforderungen am besten erfüllt, haben die Statistischen Ämter des Bundes und der Länder unterschiedliche Verfahren methodisch analysiert. Dazu haben sie Vor- und Nachteile der betrachteten Verfahren systematisiert und hinsichtlich ihrer grundsätzlichen Eignung für einen Einsatz in der Hochschulstatistik bewertet.

Demnach sind drei Arten von datenverändernden Geheimhaltungsverfahren grundsätzlich geeignet: die deterministische Rundung, Verfahren wie das zur Geheimhaltung beim Zensus 2011 eingesetzte SAFE-Verfahren[5] sowie die stochastische Überlagerung mit der Cell-Key-Methode.

Als nicht geeignet für den Einsatz in einer Auswertungsdatenbank wurde das Zellsperrverfahren bewertet, da es keine flexiblen Auswertungen ermöglicht. Das Verfahren umfasst zwei Schritte: Im ersten Schritt werden alle primär geheimzuhaltenden Tabellenfelder ermittelt. Dies sind in der Regel Zellen mit Fallzahlen von 1 oder 2. Im zweiten Schritt werden Tabellenfelder, bei denen kein primäres Enthüllungsrisiko besteht, als Sekundärsperrungen ausgewählt, um die Primärsperrungen vor Aufdeckung durch Differenzbildung zu schützen. Um eine automatisierte Geheimhaltung tabellenübergreifend durchzuführen, bedarf es eines festgelegten Tabellenkonzepts. Insofern sind keine flexiblen Auswertungen möglich.

Das Verfahren der deterministischen Rundung rundet jeden einzelnen Wert auf ein Vielfaches des Rundungswerts auf beziehungsweise ab. Bei einem Rundungswert von 3 wird beispielsweise eine Original-Fallzahl von 10

5 Die Abkürzung SAFE steht für „Sichere Anonymisierung für Einzeldaten“.

in einer geheim gehaltenen Tabelle auf 9 abgerundet und eine Original-Fallzahl von 2 auf 3 aufgerundet.

Beim SAFE-Verfahren werden in den Einzeldatensätzen Ausprägungen der Erhebungsmerkmale so verändert, dass Gruppen von völlig identischen Datensätzen entstehen. Dabei besteht jede Gruppe aus mindestens drei Datensätzen. Dieses Verfahren führt zu leicht veränderten Mikrodaten. Es wird vor der Erstellung von Ergebnistabellen angewendet.

Die stochastische Überlagerung mit der Cell-Key-Methode wird wie das Rundungsverfahren auf Auswertungstabellen angewendet. Der Überlagerungswert für ein in einer Tabelle als Randsumme dargestelltes Ergebnis errechnet sich nicht als Summe der Überlagerungswerte der entsprechenden Tabelleninnenfelder. Dieses Vorgehen vermeidet für Randwerte einen größeren Genauigkeitsverlust – so wie man auch beim Runden Tabellenrandsummen erst exakt berechnet und anschließend rundet, anstatt bereits gerundete Innenwerte aufzuaddieren. Der Hinweis „Abweichungen in den Summen ergeben sich durch Runden der Zahlen“ gilt also beim Rundungsverfahren wie auch bei stochastischer Überlagerung. Die mit diesen Verfahren behandelten Tabellen sind daher in der Regel nicht exakt additiv. Bei der Cell-Key-Methode werden Originalwerte logisch identischer Tabellenfelder – in welcher Tabelle sie auch auftreten – immer in gleicher Weise verändert. Somit führt etwa die Abfrage nach Medizin studierenden Frauen im Sommersemester 2019 an der Universität Mainz in der Auswertungsdatenbank immer zum selben Ergebnis – egal wann und in welchem Tabellenaufbau sie durchgeführt wird. Erreicht wird das, weil zunächst auf Ebene der originalen Mikrodaten für jede Einheit (Studierende, Promovierende, Prüfungen) gleichverteilte Zufallszahlen festgelegt werden. Werden für eine Auswertung die entsprechenden Datensätze zu Tabellenfeldern gruppiert, werden auch deren Zufallszahlen aufaddiert. Die Summe der Zufallszahlen, die für identische Tabellenfelder ebenfalls identisch ist, bestimmt bei der Cell-Key-Methode die Veränderung eines Tabellenfelds. Dadurch sind die Ergebnisse in den Tabellen konsistent. Das folgende Kapitel stellt die Cell-Key-Methode ausführlich dar.

Die Statistischen Ämter des Bundes und der Länder haben das SAFE-Verfahren, das Rundungsverfahren und die stochastische Überlagerung mit der Cell-Key-Methode vor dem Hintergrund der spezifischen Anforderungen der Hochschulstatistik geprüft und nach folgenden Kriterien bewertet: Schutz vor Aufdeckung, Qualität der Ergebnisse, Nutzerakzeptanz und -kommunikation sowie Flexibilität, Praktikabilität und Wirtschaftlichkeit.

Übersicht 1

Vor- und Nachteile verschiedener Geheimhaltungsverfahren

Deterministische Rundung	Stochastische Überlagerung mit der Cell-Key-Methode	SAFE-Verfahren
1. Gewährleistung der statistischen Geheimhaltung (Primärer und sekundärer Schutz)		
Alle Fallzahlen werden gerundet dargestellt	Alle ausgewiesenen Ergebnisse können verändert worden sein	Alle ausgewiesenen Ergebnisse können verändert worden sein
Relativ große Risiken, Rundungsmuster aufzubrechen	Sehr geringe Risiken, Veränderungsmuster aufzubrechen	Veränderung kann nicht aufgebrochen werden
2. Informationsverluste (Konsistenz, Additivität und Qualität)		
Ergebnisse sind durchweg konsistent, aber nicht additiv	Ergebnisse sind durchweg konsistent, aber nicht additiv	Ergebnisse sind durchweg konsistent und additiv
Durch Rundungsbasis festgelegter, geringer bis hoher Informationsverlust	Durch Parameter festgelegter, geringer Informationsverlust	Geringer Informationsverlust bei vorgeplanten Ergebnissen
3. Datennutzung (Flexibilität, Praktikabilität, Nutzerakzeptanz und -kommunikation)		
Flexible Auswertungen ohne Einschränkungen	Flexible Auswertungen ohne Einschränkungen	Auswertungen sollten vorab festgelegt werden
Dezentrale, unabhängige Anwendung möglich	Dezentrale, unabhängige Anwendung möglich	Zentrale Erzeugung der geänderten Mikrodaten, dezentrale Anwendung erst anschließend möglich
Geringer Erläuterungsbedarf	Höherer Erläuterungsbedarf	Höherer Erläuterungsbedarf
4. Wirtschaftlichkeit (Implementierungs- und regelmäßiger Aufwand)		
Geringer bis hoher Implementierungsaufwand	Hoher Implementierungsaufwand	Hoher Implementierungsaufwand
Geringer regelmäßiger Durchführungsaufwand	Geringer regelmäßiger Durchführungsaufwand (bei Implementierung in Auswertungsdatenbank)	Regelmäßig hoher maschineller Aufwand

Diese Kriterien stammen aus dem zur Auswahl eines Geheimhaltungsverfahrens entwickelten Katalog von Entscheidungskriterien (Rohde und andere, 2018). ↘ Übersicht 1

Der Vergleich der Verfahren anhand dieser Kriterien zeigt, dass alle drei Verfahren primären Schutz bieten, da alle Fallzahlen gerundet werden beziehungsweise verändert worden sein könnten. Einen hohen sekundären Schutz vor Aufdeckung bei gleichzeitig geringem Informationsverlust – auch bei tabellenübergreifenden Abgleichen – bieten jedoch nur die Cell-Key-Methode sowie das SAFE-Verfahren. Beide Verfahren weisen aufgrund geringerer Informationsverluste bei der Qualität der Ergebnisse Vorteile gegenüber dem Rundungsverfahren auf. Hinsichtlich Flexibilität und Praktikabilität weisen Rundungsverfahren und Cell-Key-Methode Vorteile gegenüber dem SAFE-Verfahren auf: Sie ermöglichen flexible Auswertungen und sind dezentral und unabhängig anwendbar. Nachteil des SAFE-Verfahrens ist der hohe regelmäßige Durchführungsaufwand. Hinsichtlich der Nutzerkommunikation und -akzeptanz ist das Rundungsverfahren einfacher zu vermitteln, bei der Cell-Key-Methode und dem SAFE-Verfahren besteht aufgrund der höheren Komplexität der Verfahren Erläuterungsbedarf.

Die Statistischen Ämter des Bundes und der Länder haben sich nach Abwägung der Vor- und Nachteile für das Verfahren der stochastischen Überlagerung mit der Cell-Key-Methode als Geheimhaltungsverfahren für die Hochschulstatistiken entschieden. Den Ausschlag dafür gaben folgende Gründe: Das Verfahren bietet durch die Überlagerung einen hohen Schutz der Originaldaten gegen Aufdeckung der Geheimhaltung bei geringstem Grad der Veränderung, ebenso lassen sich mit ihm flexibel Ergebnisse erstellen. Es ermöglicht eine dezentrale Anwendung und ist dabei mit geringem regelmäßigem Aufwand technisch implementierbar. Die Entscheidung erfolgte im Einklang mit dem Hochschulstatistikausschuss[6]. Dieser sieht zwar einen erhöhten Erklärungsbedarf der Ergebnisse wegen der (verglichen mit einem Rundungsverfahren) höheren Komplexität des Verfahrens und der gegenüber dem Rundungsverfahren stärker erklärungsbedürftigen Nicht-Additivität. Diese Nachteile werden seiner Ansicht nach aber aufgewogen durch die (bei bezüglich Aufdeckungsrisiken vergleichbarer Parametrisierung) deutlich geringeren Abweichungen der überlagerten von den Originalangaben, die die Cell-Key-Methode gegenüber dem Rundungsverfahren bewirkt.

6 Im Hochschulstatistikausschuss sind die Hauptnutzer der Hochschulstatistiken in Deutschland vertreten: die Wissenschaftsministerien von Bund und Ländern, der Wissenschaftsrat, die Hochschulen sowie Institutionen der Hochschul- und Wissenschaftsforschung.

4

Darstellung der stochastischen Überlagerung mit der Cell-Key-Methode

Das Verfahren der stochastischen Überlagerung mit der Cell-Key-Methode wurde ursprünglich vom australischen Statistikamt entwickelt (Fraser/Wooton, 2016; Thompson und andere, 2013). Damit keine Rückschlüsse auf Einzelangaben möglich sind, erhält jedes Ergebnis (hier: Tabellenfeld = Cell) im Zuge der Datenauswertung die Chance, mit einem kleinen „Überlagerungswert“ verändert zu werden. Anstelle des Originalergebnisses wird jeweils die Summe aus Originalergebnis und Überlagerungswert veröffentlicht. Man spricht von einer Überlagerung der Originalergebnisse. Der Ermittlung eines Überlagerungswerts liegt eine einmalig festzulegende (Wahrscheinlichkeits-)Verteilung mit möglichen Überlagerungswerten zugrunde. Ein deterministischer Mechanismus sorgt dabei in Kombination mit Zellhäufigkeit (beziehungsweise Fallzahl) und Zellschlüssel (Cell-Key) dafür, dass aus der Verteilung ein eindeutiger Überlagerungswert gezogen wird. Eine mit diesem Vorgehen geheim gehaltene Tabelle erfüllt die Anforderung der tabellenübergreifenden Konsistenz, führt aber nicht zwangsläufig zu additiven Ergebnissen. Auf die beiden Eigenschaften Konsistenz und Nicht-Additivität wird in Abschnitt 4.3 näher eingegangen.

Die wichtigsten Komponenten des Verfahrens sind:

› Cell-Key-Bestimmung (Abschnitt 4.1): Ein Tabellierungswerkzeug oder eine Auswertungsdatenbank muss im Zuge der Tabellenerstellung parallel zur Bestimmung der Häufigkeit einen Cell-Key berechnen.

› Übergangsmatrix (Abschnitt 4.2): Die statistischen Eigenschaften der Überlagerung werden in Form einer Übergangsmatrix festgelegt. Diese enthält die beding-

ten Übergangswahrscheinlichkeiten[7] und beschreibt die statistikspezifischen Regeln des deterministischen Überlagerungsprozesses.[8]

› Lookup-Modul (Abschnitt 4.3): Der eigentliche Überlagerungsprozess wird durch den sogenannten Lookup-Schritt umgesetzt. Anhand des Wertepaares – bestehend aus Zellhäufigkeit und Cell-Key – wird in einem deterministischen Schritt der Überlagerungswert aus der Überlagerungsmatrix abgelesen.

4.1 Bestimmung der Cell-Keys

Der Prozess zur Bestimmung der Cell-Keys muss in dem Sinne konsistent sein, dass er inhaltlich und logisch identischen Tabellenfeldern – und zwar unabhängig von der jeweils betrachteten Tabelle – ein und denselben Cell-Key zuweist.

Bei der stochastischen Überlagerung mit der Cell-Key-Methode handelt es sich um ein Geheimhaltungsverfahren, bei dem Ergebnisse erst im Zuge der Datenauswertung verändert werden. Dennoch wird bereits im Originaldatenbestand (in der Regel auf Mikrodatenebene) jeder Beobachtungseinheit (Record) einmalig eine gleichverteilte Zufallszahl zugewiesen. Dieser sogenannte Record-Key wird in der nachfolgend dargestellten Umsetzungsvariante aus einer Gleichverteilung im Intervall [0, 1) gezogen. Die Zufallsziehung der Record-Keys stellt die eigentliche stochastische Komponente des Verfahrens dar.

Bei der Berechnung von Tabellenergebnissen entsprechend den Merkmalsgliederungen sind neben der standardmäßigen Aggregation von Beobachtungen zu Häufigkeiten auch die Summen der Record-Keys eines Tabellenfelds zu bilden. Die Nachkommastellen der dabei aufsummierten Record-Keys ergeben dann die Cell-Keys, die wie die Record-Keys im Intervall [0, 1) gleichverteilt sind (Tent, 2019).

↘ Wie wird ein Cell-Key berechnet?

Angenommen, ein Mikrodatensatz mit zehn Beobachtungen enthält drei Professorinnen. Die zuvor gezogenen Record-Keys der drei Beobachtungen betragen 0,6019, 0,8531 und 0,3448. Wird das Einzelmaterial nach dem Merkmal Geschlecht tabelliert dargestellt, so ergibt sich für „Professorinnen“ eine Zellhäufigkeit von 1 + 1 + 1 = 3 sowie der entsprechende Cell-Key von 0,7998 (nach Aufsummieren der Record-Keys der zur Zellhäufigkeit beitragenden Beobachtungseinheiten 0,6019 + 0,8531 + 0,3448 = 1,7998 und ausschließlicher Betrachtung der Nachkommastellen).

4.2 Übergangsmatrix – Design der stochastischen Eigenschaften

Bei dezentral organisierten Bundesstatistiken werden die stochastischen Eigenschaften der Überlagerungen einmalig und einheitlich für eine Statistik festgelegt. Dies erfolgt im Verbund der amtlichen Statistik durch das fachlich zuständige Gremium. Zu den wichtigsten Eigenschaften der stochastischen Überlagerung gehören:

› Unverzerrtheit der Überlagerungen: Der Überlagerungswert, der zu den Originalergebnissen addiert wird, nimmt im Mittel den Wert 0 an.

› Konstante Streuung der Verteilung der Überlagerungen.

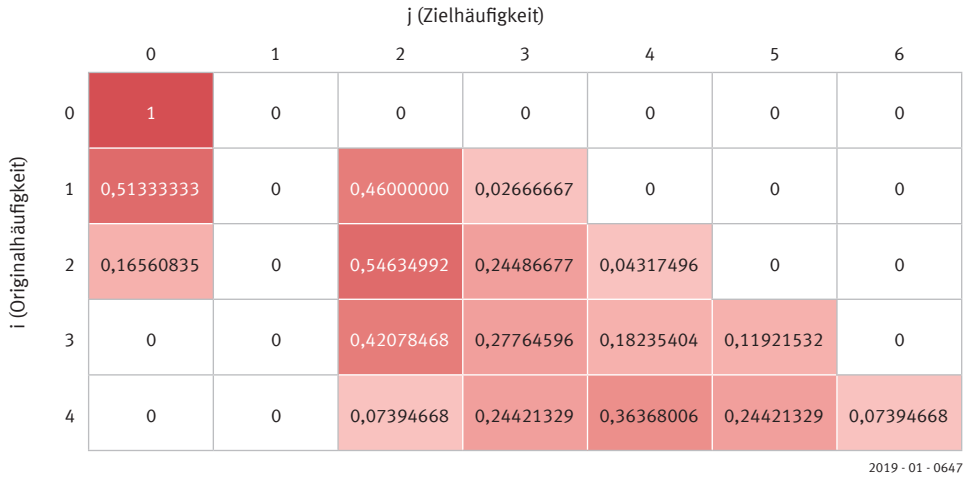
Zur Umsetzung der stochastischen Eigenschaften der Überlagerung werden Wahrscheinlichkeitsverteilungen für die Überlagerungen festgelegt. Es handelt sich dabei um als sogenannte Übergangsmatrix notierte bedingte Wahrscheinlichkeitsverteilungen. Bedingt bedeutet in diesem Zusammenhang bedingt auf feste Originalhäufigkeiten. Bedingte Wahrscheinlichkeitsverteilungen werden benötigt, da je nach Originalhäufigkeit *i* die sinnvollen Zielhäufigkeiten *j* abweichen können. So sollen Originalhäufigkeiten von 0 nicht verändert werden und in den veröffentlichten Ergebnissen keine negativen Werte enthalten sein. Die zu einer Originalhäufigkeit *i* korrespondierende Zeile der Übergangsmatrix legt fest, welche Wahrscheinlichkeit die Überlagerung dieser Originalhäufigkeit *i* hin zur Zielhäufigkeit *j* haben soll[9].

7 Man spricht von einer bedingten (Übergangs-)Wahrscheinlichkeit, da die Wahrscheinlichkeit für das Eintreten einer bestimmten Überlagerung in Abhängigkeit von der Größe der Originalhäufigkeit festgelegt wird.

8 Bei der Implementierung des Verfahrens wird die Übergangsmatrix in eine für IT-Umsetzungen geeignete Form transponiert. Man spricht dann von einer Überlagerungsmatrix beziehungsweise -tabelle.

9 Lesehilfe: Diese Zeile definiert die Wahrscheinlichkeitsverteilung der Zielhäufigkeiten unter der Bedingung einer Originalhäufigkeit von *i*.

Grafik 1
Beispiel für eine Übergangsmatrix



↘ Grafik 1 zeigt ein anschauliches Beispiel für eine mögliche Realisierung einer Übergangsmatrix. In diesem Beispiel wurde die Eigenschaft gesetzt, dass Werte von 1 nicht im geheim gehaltenen Ergebnis enthalten sind (siehe auch Erläuterung im weiteren Abschnitt).

Die Eigenschaften und somit die konkrete Ausgestaltung der Übergangsmatrix können anhand von Verfahrensparametern und weiteren Vorgaben unmittelbar gesteuert werden. Relevante Verfahrensparameter zur Bestimmung der Übergangsmatrix sind dabei:

› die Maximalabweichung, das heißt der Betrag der maximalen Abweichung zwischen Originalhäufigkeit und Zielhäufigkeit,

› die Varianz, das heißt das Streuungsmaß der Verteilung der Abweichungen.

Zudem kann eine Bleibewahrscheinlichkeit vorgegeben werden (in Grafik 1 sind dies Diagonaleinträge in der Übergangsmatrix). Diese gibt an, mit welcher Wahrscheinlichkeit eine Originalhäufigkeit i unverändert bleibt und somit nicht überlagert wird. Des Weiteren kann die Eigenschaft gesetzt werden, dass bestimmte Häufigkeiten nicht im geheim gehaltenen Ergebnis enthalten sind.[10]

↘ Wie wird eine Überlagerungsmatrix anhand der vorgegebenen Eigenschaften und Parameter berechnet?

Um die Übergangsmatrix bestimmen zu können, wird für jede Originalhäufigkeit $i = (0, 1, 2, ..., L)$ eine bedingte Wahrscheinlichkeitsverteilung p_i mit den Wahrscheinlichkeiten für die Übergänge v_i hin zu den Zielhäufigkeiten j gesucht. Die Wahrscheinlichkeiten ergeben sich als Lösungen nicht-linearer Gleichungssysteme, die durch die folgenden Eigenschaften in Form von Nebenbedingungen und Restriktionen aufgespannt werden (Giessing, 2016):

(1) Unverzerrtheit der Überlagerungen: $p_i v_i = 0$, wobei v_i die Überlagerungen darstellen beziehungsweise $\sum_j p_{ij} v_{ij} = 0$

(2) Konstante Varianz[11]: $p_i(v_i)^2 = Var$ beziehungsweise $\sum_j p_{ij}(v_{ij})^2 = Var$

(3) Die Überlagerungen führen zu keinen negativen Zielhäufigkeiten. Zudem sollen keine von null verschiedenen Zielhäufigkeiten kleiner gleich eines vorzugebenden Schwellenwertes j_s, $j_s \geq 0$ erreicht werden.

(4) Eine Originalhäufigkeit kann maximal mit einem Wert D überlagert werden.

(5) Die Übergangswahrscheinlichkeiten einer Originalhäufigkeit summieren sich zu eins.

10 Im kommenden Zensus 2021 sollen Werte von 1 und 2 nicht in den veröffentlichten Ergebnissen dargestellt werden. Das Beispiel in Grafik 1 hingegen passt zu einem anderen Szenario, in dem Werte von 1 nicht enthalten wären.

11 Die Darstellung der Varianz berücksichtigt die wegen (1) gegebene Unverzerrtheit der Überlagerungen.

(6) Die einzelnen Übergangswahrscheinlichkeiten sind positiv, jedoch kleiner eins.

Die Suche nach einem geeigneten Kandidaten p_i^* unter Einhaltung aller sechs Eigenschaften reduziert sich zur Lösung eines linearen Gleichungssystems der Form $Ap_i - b = 0$. In bestimmten Konstellationen kann es vorkommen, dass es keine oder keine eindeutige Lösung gibt (unter anderem, wenn mehr Übergänge als Kriterien vorliegen; Gießing/Höhne, 2010). Es bedarf daher einer Vorgehensweise, um aus der möglichen Vielzahl an Lösungen eine Wahrscheinlichkeitsverteilung den anderen vorzuziehen. Eine in Marley und Leaver (2011) vorgeschlagene Möglichkeit, eine optimale Lösung zu bestimmen, ist die Maximum-Entropy-Methode aus der Informationstheorie. Sie beruht auf dem Prinzip, das von Reiter (1985) wie folgt beschrieben wird:

«Ist auf der Grundlage unzureichender Information aus einer Vielzahl von Wahrscheinlichkeitsverteilungen eine Verteilung auszuwählen, dann ist genau diejenige zu nehmen, welche die größte Entropie besitzt und mit der gesamten verfügbaren Information übereinstimmt.»

Die zu maximierende Entropie lautet:

$$S(p_i) := -\sum_j p_{ij} \log_2(p_{ij})$$

Das Ziel dieses Vorgehens ist es, eine Übergangsmatrix zu bestimmen, die den mittleren Informationsgehalt (das heißt den Erwartungswert des Logarithmus der Wahrscheinlichkeiten) maximiert. |12

4.3 Der deterministische Lookup-Schritt – Implementierung

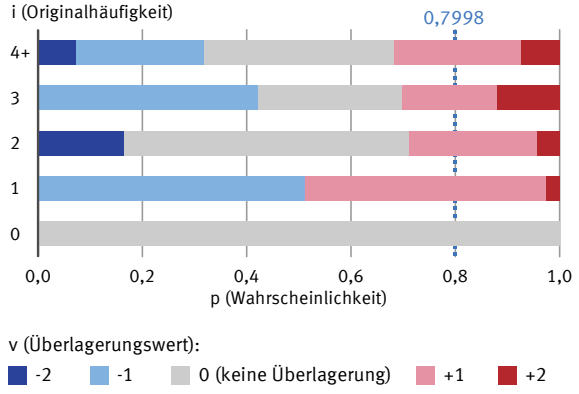
Im letzten Verfahrensschritt, dem sogenannten Lookup, wird für jedes Tabellenfeld anhand der Originalhäufigkeit und des dazugehörigen Cell-Keys der jeweilige Überlagerungswert abgelesen. Der Überlagerungswert wird zum jeweiligen Originalergebnis hinzuaddiert und stellt die Differenz aus Ziel- und Originalhäufigkeit dar. Wegen des konsistenten Cell-Keys für logisch identische Ausprägungskombinationen ist er immer identisch. Auf diese Weise liefert das Verfahren konsistente Tabellen und muss dazu grundsätzlich alle Ergebnisse gleichbehandeln – auch Rand- und Zwischensummen.

12 Für die Berechnung einer Übergangsmatrix wird das für die Statistiksoftware R entwickelte „ptable"-Paket (Enderle, 2019) verwendet.

↘ Wie wird ein Überlagerungswert anhand von Zellhäufigkeit und Cell-Key ermittelt? (Fortsetzung des Beispiels)

Um die Originalhäufigkeit 3 (der drei Professorinnen) zu überlagern, wird im entsprechenden Balken des Überlagerungstableaus |13 der Überlagerungswert an der Stelle 0,7998 abgelesen. Es ergibt sich ein Überlagerungswert von +1 (die 0,7998 schneidet den Balken im hellroten Bereich). Aus der originalen Zellhäufigkeit 3 wird nach Überlagerung eine zu veröffentlichende 4. ↘ Grafik 2

Grafik 2
Überlagerungstableau



2019 - 01 - 0648

Aufgrund der Gleichbehandlung aller Tabellenfelder und dem damit einhergehenden Genauigkeitsvorteil (siehe Kapitel 3) sind die mit diesem Geheimhaltungsverfahren behandelten Tabellen in der Regel nicht exakt additiv (das heißt mathematisch betrachtet müssen die linearen Beziehungen in einer Tabelle nicht erfüllt sein).

13 Die Übergangsmatrix (Beispiel siehe Grafik 1) kann grafisch als ein sogenanntes Überlagerungstableau dargestellt werden. Jeder Balken entspricht einem Originalwert. Unterschiedliche Farben entsprechen unterschiedlichen Überlagerungen und die Breite des farbigen Teilbalkens entspricht der in der Übergangsmatrix vorgegebenen Wahrscheinlichkeit, mit der es zu der betreffenden Überlagerung des jeweiligen Originalwerts kommt. Der Lookup-Schritt „liest" die Überlagerung im Überlagerungstableau in der durch den Originalwert *i* gegebenen Zeile an der Stelle p = cell key ab.

Warum wird die Nicht-Additivität in Kauf genommen?

Durch das unabhängige und separate Überlagern von Tabellenfeldern sind zwei wichtige Vorteile gegeben:

(1) Tabellenübergreifende Konsistenz
Der zur Originalhäufigkeit hinzuzuaddierende Überlagerungswert eines bestimmten Ergebnisses (Beispiel: Anzahl der Studierenden im Studienfach Mathematik) ist aufgrund des Cell-Key-Vorgehens und des deterministischen Lookup-Schrittes unabhängig von der Tabelle, in der das Ergebnis dargestellt wird, immer identisch. Dabei macht es keinen Unterschied, ob es sich in einer Tabellendarstellung um eine Randsumme der beiden Innenfelder (hier: „Studierende im Studienfach Mathematik, männlich“ und „Studierende im Studienfach Mathematik, weiblich“) handelt oder in einer weiteren um eine Zusammenfassung nach Ländern (zum Beispiel „Deutsche Studierende im Studienfach Mathematik“ und „Ausländische Studierende …“).

(2) Genauigkeit
Es wird vermieden, dass sich eine Reihe zufällig gleichgerichteter Überlagerungen in Summen kumulieren und dann im Einzelfall etwas größere Veränderungen zwischen Original- und geheim gehaltenen Werten hervorrufen. Um beim obigen Beispiel zu bleiben: Die „Anzahl an Studierenden im Studienfach Mathematik“ wird in einer Tabelle als Summe der entsprechenden Studierenden aller einzelnen Hochschulen dargestellt und in allen Summanden fällt der Überlagerungswert zufällig negativ aus. In diesem Fall würde ein als Summe aller Einzelüberlagerungen gebildeter, nicht mehr ganz so kleiner Überlagerungswert das Gesamtergebnis unnötig „kräftig“ verkleinern.

5

Entscheidungsprozess zur Auswahl statistikspezifischer Parameter für die Hochschulstatistiken

Um die für den Einsatz in der Hochschulstatistik geeigneten Parameter zu finden, wurde eine zweistufige Evaluationsstudie durchgeführt und mithilfe sogenannter Monte-Carlo-Simulationsstudien die statistikspezifischen Aspekte der Hochschulstatistik nachgestaltet.

In der ersten Stufe wurden Untersuchungen zu Varianten der Cell-Key-Methode mit jeweils unterschiedlichen Parametervorgaben (unter anderem Maximalabweichung) und vorgegebenen Bleibewahrscheinlichkeiten auf Basis ausgewählter fester Auswertungstabellen der Hochschulstatistik durchgeführt. Als Entscheidungskriterien (Rohde und andere, 2018) dienten Indikatoren zur Qualität (unter anderem Informationsverlustmaße wie der mittlere empirische Betrag der Abweichungen) und zur Schutzwirkung (insbesondere zum sekundären Aufdeckungsrisiko; Enderle und andere, 2018).

In der zweiten Stufe wurden diese Untersuchungen auf Basis flexibler Auswertungstabellen durchgeführt. Bei flexibler Auswertung (beispielsweise über eine dynamische Auswertungsdatenbank) sind im Vergleich zur Untersuchung fester Auswertungstabellen (wie in klassischen Printmedien) die Aufdeckungsrisiken leicht erhöht. Dem kann aber durch geeignete Parametrisierung der stochastischen Überlagerung gut begegnet werden.

Aufgrund der Entscheidung, Verhältniszahlen – wie etwa Studierenden-/Personalrelationen – auf der Basis bereits überlagerter Fallzahlen zu berechnen, wurden die Auswirkungen unterschiedlicher Parametrisierungen auch im Hinblick auf die Qualität von Verhältniszahlen betrachtet. Hierfür durchgeführte Berechnungen zeigten, dass Unterschiede zwischen den für die Hochschulstatistik in Betracht gezogenen Parametrisierungen der stochastischen Überlagerung keinen relevanten Einfluss auf die Qualität der Verhältniszahlen haben. Die Qualität der auf der Basis von überlagerten Fallzahlen berechneten Verhältniszahlen wird im Wesentlichen nicht durch Bleibewahrscheinlichkeiten oder Maximalabweichungen bestimmt. Vielmehr beeinflusst in erster Linie die Größe der in die Berechnung von Verhältniszahlen eingehenden Fallzahlen deren Qualität. Generell beeinträchtigen kleine Fallzahlen die Qualität und somit die Aussagekraft von Verhältniszahlen. Eine hinreichende statistische Aussagefähigkeit ergibt sich erst ab einer bestimmten Größe der dahinterstehenden Basiszahlen in Zähler und Nenner.

Als Ergebnis der Evaluationsstudien wurde statistikintern eine Parametrisierung festgelegt, die den nachfolgenden Qualitätskriterien genügt. Die Qualitätskriterien werden künftig mit den hochschulstatistischen Ergebnissen veröffentlicht, um über das Ausmaß der Veränderung der hochschulstatistischen Ergebnisse durch die Cell-Key-Methode zu informieren:

1. Der in den Tabellen üblicherweise zu erwartende mittlere Betrag der Abweichung zwischen überlagerten und originalen Fallzahlen liegt unter 0,5.
2. Mindestens 90 % der Fallzahlen in den Tabellen bleiben unverändert oder weichen um maximal 1 vom Originalwert ab.
3. Bei höchstens 5 % der Fallzahlen in den Tabellen liegt die Abweichung bei 3 oder mehr.
4. Bei höchstens 0,5 % der Fallzahlen in den Tabellen liegt die Abweichung bei 4 oder mehr.

Ein Kommunikationskonzept regelt zudem, wie den Nutzerinnen und Nutzern der Auswertungsdatenbank erforderliche Interpretationshilfen für selbstberechnete Verhältniszahlen zur Verfügung gestellt werden können.

6

Fazit und Ausblick

Die Geheimhaltung in der Hochschulstatistik erfolgt künftig durch die in diesem Aufsatz beschriebene stochastische Überlagerung mit der Cell-Key-Methode. Die geplante Auswertungsdatenbank Hochschulstatistik wird mit ihrer Inbetriebnahme flexible Auswertungen auf Basis der eingespeicherten Einzelangaben der Hochschulstatistiken ermöglichen. Die in die Datenbank integrierte Cell-Key-Methode sorgt dann für eine vollautomatisierte und konsistente Geheimhaltung aller Ergebnistabellen. Nach der Festlegung spezifischer Parameter für die Cell-Key-Methode durch die Statistischen Ämter des Bundes und der Länder wird derzeit die Implementierung der Methode in die Auswertungsdatenbank vorbereitet. Außerdem wird eine Nutzerdokumentation zur Cell-Key-Methode entwickelt, die das Verfahren erläutert, das Ausmaß der Veränderungen umreißt und so die Interpretation der über die Cell-Key-Methode veränderten Ergebnisse unterstützt.

Vorgesehener Produktivstart der Auswertungsdatenbank Hochschulstatistik ist Anfang 2021. Zunächst werden nur Mitarbeiterinnen und Mitarbeiter der statistischen Ämter einen Zugang zur Datenbank erhalten, in einer Ausbaustufe ist auch ein Datenzugang für externe Nutzerinnen und Nutzer vorgesehen.

LITERATURVERZEICHNIS

Enderle, Tobias. *ptable: A Perturbation Table Generator for SDC Tools*. github.com/sdcTools/ptable.

Enderle, Tobias/Giessing, Sarah/Tent, Reinhard. *Designing Confidentiality on the fly Methodology – Three Aspects*. In: Domingo-Ferrer, Josep/Montes, Francisco (Herausgeber). Privacy in Statistical Databases. LNCS (Lecture Notes in Computer Science). 2018. Ausgabe 11126, Seite 28 ff. [Zugriff am 1. November 2019]. Verfügbar unter: doi.org/10.1007/978-3-319-99771-1_3

Fraser, Bruce/Wooton, Janice. *A proposed method for confidentialising tabular output to protect against differencing*. Work session on Statistical Data Confidentiality. Supporting paper. Genf 2005. [Zugriff am 1. November 2019]. Verfügbar unter: www.unece.org

Giessing, Sarah. *Computational Issues in the Design of Transition Probabilities and Disclosure Risk Estimation for Additive Noise*. In: Domingo-Ferrer, Josep/Peji-Bach, Mirjana (Herausgeber). Privacy in Statistical Databases. LNCS (Lecture Notes in Computer Science). 2016. Ausgabe 9867, Seite 237 ff.

Giessing, Sarah/Höhne, Jörg. *Eliminating Small Cells from Census Counts Tables: Some Considerations on Transition Probabilities*. In: Domingo-Ferrer, Josep/Magkos, Emmanuel (Herausgeber). Privacy in Statistical Databases. LNCS (Lecture Notes in Computer Science). 2010. Ausgabe 6344, Seite 52 ff.

Marley, Jennifer K./Leaver, Victoria L. *A method for confidentialising user-defined tables: Statistical properties and a risk-utility analysis*. In: Proceedings of 58th World Statistical Congress. 2011, Seite 1072 ff.

Reiter, Johann. *Bildverschärfung durch Lösung der Fredholmschen Integralgleichung 1. Art mittels der Maximum-Entropie-Methode mit astronomischen Anwendungen*. Dissertation. Innsbruck 1985.

Rohde, Johannes/Seifert, Christian/Gießing, Sarah. *Entscheidungskriterien für die Auswahl eines Geheimhaltungsverfahrens*. In: WISTA Wirtschaft und Statistik. Ausgabe 3/2018, Seite 90 ff.

Tent, Reinhard. *Cell-Keys – Ein mathematischer Beweis für die Gleichverteilung*. Internes Arbeitspapier des Statistischen Bundesamtes. 2019.

Thompson, Gwenda/Broadfoot, Stephen/Elazar, Daniel. *Methodology for the Automatic Confidentialisation of Statistical Outputs from Remote Servers at the Australian Bureau of Statistics*. Paper presented at the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality. Ottawa 2013. [Zugriff am 1. November 2019]. Verfügbar unter: www.unece.org

RECHTSGRUNDLAGEN

Gesetz über die Statistik für das Hochschulwesen sowie für die Berufsakademien (Hochschulstatistikgesetz – HStatG) vom 2. November 1990 (BGBl. I Seite 2414), das zuletzt durch Artikel 3 des Gesetzes vom 7. Dezember 2016 (BGBl. I Seite 2826) geändert worden ist.

Gesetz über die Statistik für Bundeszwecke (Bundesstatistikgesetz – BStatG) in der Fassung der Bekanntmachung vom 20. Oktober 2016 (BGBl. I Seite 2394), das zuletzt durch Artikel 10 Absatz 5 des Gesetzes vom 30. Oktober 2017 (BGBl. I Seite 3618) geändert worden ist.

Herausgeber
Statistisches Bundesamt (Destatis), Wiesbaden

Schriftleitung
Dr. Daniel Vorgrimler
Redaktionsleitung: Juliane Gude
Redaktion: Ellen Römer

Ihr Kontakt zu uns
www.destatis.de/kontakt

Erscheinungsfolge
zweimonatlich, erschienen im Dezember 2019
Das Archiv älterer Ausgaben finden Sie unter www.destatis.de

Print
Einzelpreis: EUR 19,– (zzgl. Versand)
Jahresbezugspreis: EUR 114,– (zzgl. Versand)
Bestellnummer: 1010200-19006-1
ISSN 0043-6143
ISBN 978-3-8246-1085-3

Download (PDF)
Artikelnummer: 1010200-19006-4, ISSN 1619-2907

Vertriebspartner
IBRo Versandservice GmbH
Bereich Statistisches Bundesamt
Kastanienweg 1
D-18184 Roggentin
Telefon: +49 (0) 382 04 / 6 65 43
Telefax: +49 (0) 382 04 / 6 69 19
destatis@ibro.de

Papier: Design Offset, FSC-zertifiziert