DIE CELL-KEY-METHODE

in den Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder

Teil 1: Vorstellung des neuen Geheimhaltungsverfahrens

Stefanie Setzer, Johannes Rohde, Volker Güttgemanns, Patrick Rothe

Die Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder führen für ausgewählte Statistiken die Cell-Key-Methode als neues Verfahren zur Ergebnisgeheimhaltung ein. Dieses Verfahren schützt die Befragten vor der Reidentifikation, indem es durch die Überlagerung der Fallzahlen eine Unsicherheit über die Anzahl der tatsächlich zum Ergebnis beitragenden Fälle schafft. Der Artikel stellt die Funktionsweise der Cell-Key-Methode vor und bietet dabei sowohl eine einfach zu verstehende Einführung in die Thematik als auch detaillierte methodische Informationen.

terovesalainen / stock.adobe.com

1 Einleitung

Die Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder (FDZ) stellen Mikrodaten für die wissenschaftliche Nutzung bereit. Um den Datenschutz hierbei zu gewährleisten, gibt es zwei Möglichkeiten: entweder die Anonymisierung, bei der die Daten vor der Bereitstellung an die Wissenschaft so verändert werden, dass bei der Auswertung keine Geheimhaltungsrisiken entstehen können, oder die Geheimhaltung, bei welcher der Schutz der Daten durch eine Veränderung der Ergebnisse erzeugt wird. Die Wahl der Vorgehensweise hängt dabei vom gewählten Zugangsweg ab: Bei den sogenannten Off-Site-Zugangswegen, bei denen die Daten an die Nutzenden übermittelt werden, erfolgt eine Anonymisierung der Daten, was jedoch immer mit einem Informationsverlust einhergeht. Bei Nutzung der On-Site-Zugangswege, also eines Gastwissenschaftsarbeitsplatzes oder der kontrollierten Datenfernverarbeitung, verbleiben die Daten in den geschützten Räumen der amtlichen Statistik. Dort kann in der Regel das volle Informationspotenzial der Daten erhalten bleiben, die erzeugten Ergebnisse werden dafür aber einer Geheimhaltungsprüfung unterzogen.

Die Auswirkungen dieser Geheimhaltungsprüfung kennt jede Person, die schon einmal einen der On-Site-Zugangswege der Forschungsdatenzentren genutzt hat: Nach Bereitstellung der Ergebnisse springen häufig drei große X ins Auge. Dieses Sperrmuster verwenden die Forschungsdatenzentren üblicherweise, wenn die Veröffentlichung eines Ergebnisses ein Geheimhaltungsrisiko darstellt. Doch warum nehmen die Forschungsdatenzentren die Geheimhaltung überhaupt so ernst? Gäbe es Alternativen zu den drei großen X?

Kapitel 2 beantwortet zunächst die erste Frage, warum die Geheimhaltung den Forschungsdatenzentren so wichtig ist. Wie die Cell-Key-Methode als alternatives Verfahren zur Sperrung mit den drei großen X funktioniert, erläutert Kapitel 3. Danach erläutert Kapitel 4 die Methodik der Cell-Key-Methode formell. Ein kurzes Fazit mit dem Hinweis auf den zweiten Aufsatzteil beschließt den Beitrag.

Stefanie Setzer
ist Diplom-Soziologin und Referentin im Referat „Forschungsdatenzentrum, Methoden der Datenanalyse" des Statistischen Bundesamtes. Schwerpunkt ihrer Arbeit ist die fachliche und methodische Weiterentwicklung des Arbeitsbereichs.

Dr. Johannes Rohde
hat Wirtschaftswissenschaften an der Leibniz Universität Hannover studiert und dort 2015 seine Promotion im Bereich Statistik abgeschlossen. Bei IT.NRW leitet er den Service „Mathematisch-statistische Methoden und experimentelle Statistik".

Volker Güttgemanns
hat einen Master of Science in Wirtschaftswissenschaften und war von 2017 bis 2023 stellvertretende Leitung der Geschäftsstelle des Forschungsdatenzentrums der Statistischen Ämter der Länder.

Patrick Rothe
hat Sozialwissenschaften an der Universität Mannheim studiert und ist seit 2011 im Bayerischen Landesamt für Statistik tätig. Seit 2018 leitet er dort das Sachgebiet „Grundsatzfragen der amtlichen Statistik, Digitalisierung, Forschungsdatenzentrum, Kompetenzzentrum Analyse". Inhaltlich beschäftigt er sich schwerpunktmäßig unter anderem mit der statistischen Geheimhaltung.

Der vorliegende Beitrag ist in der Zeitschrift WISTA Wirtschaft und Statistik, Ausgabe 3/2024 erschienen und wird hier im Originalwortlaut mit Originalabbildungen abgedruckt. Das Bayerische Landesamt für Statistik dankt den Autoren und dem Statistischen Bundesamt (Destatis) für die freundliche Nachdruckgenehmigung.

> *Der Schutz der anvertrauten Daten hat daher für die amtliche Statistik – und damit auch für die Forschungsdatenzentren – stets die oberste Priorität.*

2 Geheimhaltung in den Forschungsdatenzentren

2.1 Warum nehmen die Forschungsdatenzentren Geheimhaltung so ernst?

Diese Frage lässt sich einfach beantworten: weil sie gesetzlich dazu verpflichtet sind. Die Pflicht zur Geheimhaltung ist in § 16 Bundesstatistikgesetz (BStatG) geregelt. Danach sind „Einzelangaben über persönliche und sachliche Verhältnisse, die für eine Bundesstatistik gemacht werden, […] geheim zu halten" (§ 16 Absatz 1 BStatG). Dieser Absatz regelt aber auch Ausnahmen, die es den statistischen Ämtern und der Wissenschaft ermöglichen, Daten und Ergebnisse unter bestimmten Voraussetzungen zu veröffentlichen. Eine Veröffentlichung ist beispielsweise möglich, wenn die Einzelangaben mit den Ergebnissen anderer Befragter zusammengefasst wurden oder wenn die Einzelangaben den Betroffenen nicht zuzuordnen sind. Diese beiden Ausnahmen ermöglichen die Bereitstellung von Daten und Ergebnissen und begründen gleichzeitig die Pflicht zur Geheimhaltung. Ergebnisveröffentlichungen sind erlaubt, solange aus den Ergebnissen keine Rückschlüsse auf Einzelne gezogen werden können.

Der Grund für diese gesetzliche Regelung und den dadurch festgelegten hohen Stellenwert der Geheimhaltung ist gut nachvollziehbar: Für die Erhebungen der amtlichen Statistik besteht oft Auskunftspflicht. Die Befragten – seien es Personen, Unternehmen, Betriebe oder Sonstige – können demnach häufig nicht selbst entscheiden, welche Informationen sie von sich preisgeben wollen. Um diesen Eingriff in die informationelle Selbstbestimmung auszugleichen, garantiert der Gesetzgeber den Befragten, dass ihnen ihre Angaben nicht zugeordnet werden können. Gleiches gilt für Erhebungen mit freiwilliger Teilnahme.

Das Vertrauen der Befragten in die Nicht-Zuordenbarkeit ihrer Angaben ist die Grundlage dafür, dass Fragen ohne Sorge vor Enthüllung wahrheitsgemäß beantwortet werden, und trägt somit maßgeblich zur hohen Qualität der Daten bei. Der Schutz der anvertrauten Daten hat daher für die amtliche Statistik – und damit auch für die Forschungsdatenzentren – stets die oberste Priorität.

2.2 Der bisherige Standard: die Zellsperrung

Bisher stellen die statistischen Ämter die Geheimhaltung in der Regel mithilfe der Zellsperrung sicher.[1] Bei dieser Form der Geheimhaltung werden alle Angaben, die ein Geheimhaltungsrisiko darstellen, durch ein Sperrmuster („XXX") ersetzt. Dieses Verfahren hat sich in der amtlichen Statistik bewährt, weist jedoch einige gravierende Nachteile auf:

- **Informationsverlust:** Bei der Zellsperrung werden nicht nur die kritischen Angaben selbst gesperrt (Primärsperrung). Um eine Rückrechnung dieser Werte zu verhindern, müssen sie mit an sich unkritischen Angaben gegengesperrt werden (Sekundärsperrung). Wenn gesperrte Angaben über andere Tabellen rückrechenbar sind, müssen auch hier Sperrungen umgesetzt werden (tabellenübergreifende Sperrung). So kann ein einzelner zu sperrender Wert schnell eine Vielzahl weiterer Sperrungen an sich unkritischer Werte nach sich ziehen.

- **Hoher Aufwand:** Da die Geheimhaltungsprüfung bei der Zellsperrung in der Regel nicht vollständig automatisiert erfolgen kann, ist die Geheimhaltung sehr zeit- und ressourcenintensiv und Nutzende müssen teils lange auf ihre Ergebnisse warten.

- **Unzufriedenheit:** Durch dieses Verfahren können nicht alle interessierenden Werte veröffentlicht werden, teilweise müssen sogar ganze Tabellen gesperrt werden. Daher führt die Zellsperrung häufig zu unzufriedenen Datennutzenden.

1 Beim Zensus 2011 wurde außerdem das Verfahren SAFE – Sichere Anonymisierung Für Einzelangaben genutzt (Höhne, 2015).

2 Die Cell-Key-Methode wurde ursprünglich vom australischen Statistikamt (Australian Bureau of Statistics) entwickelt.

2.3 Das neue Geheimhaltungsverfahren: die Cell-Key-Methode

Aufgrund der beschriebenen Nachteile der Zellsperrung haben die statistischen Ämter für erste Statistiken die Einführung der Cell-Key-Methode (CKM) beschlossen.[2] Diese Entscheidung wirkt sich unmittelbar auf die Datenbereitstellung in den Forschungsdatenzentren aus, da die Sicherstellung der Geheimhaltung stets in Einklang mit den fachseitig festgelegten Geheimhaltungsregeln erfolgt. Bei der Cell-Key-Methode handelt es sich um ein datenveränderndes Verfahren für die Geheimhaltung von Fallzahltabellen. Die Vorteile des Verfahrens sind:

- Mit der Cell-Key-Methode gibt es keine Sperrungen.

- Die Ergebnisse weisen eine hohe Datenqualität auf und sind tabellenübergreifend konsistent.

- Der Aufwand für die Geheimhaltungsprüfung von Tabellen ist deutlich geringer.

- Aufdeckungsrisiken durch Fehler bei der tabellenübergreifenden Geheimhaltung können ausgeschlossen werden.

Diesen Vorteilen stehen aber auch Nachteile gegenüber:

- Tabellen sind nach der Anwendung der Cell-Key-Methode nicht mehr additiv.

- Gerade bei kleinen Fallzahlen kann die Veränderung der Werte relativ stark ausfallen.

- Da es sich bei der Cell-Key-Methode originär um ein Verfahren für die Geheimhaltung von Fallzahltabellen handelt, können zusätzliche Aufwände bei der Prüfung der Ergebnisse multivariater Analysemethoden entstehen.

- Die Cell-Key-Methode ist weniger eingängig als andere Geheimhaltungsverfahren und bedarf daher umfangreicher Erläuterung.

Das Verfahren der Cell-Key-Methode wird im Folgenden für die Grundform der Geheimhaltung von Fallzahltabellen vorgestellt.

3 Funktionsweise der Cell-Key-Methode

Bei der Cell-Key-Methode handelt es sich um ein posttabulares datenveränderndes Geheimhaltungsverfahren. Das bedeutet, dass das Verfahren erst bei der Ergebniserstellung ansetzt und dass die Geheimhaltung durch eine Veränderung von Fallzahlen erfolgt. Die Schutzwirkung wird dadurch erzielt, dass Unsicherheit bezüglich der Originalfallzahl geschaffen wird, indem Tabellenwerte mit einem Fehlerterm überlagert werden. Das Verfahren stellt dabei sicher, dass die Überlagerung konsistent ist, dass also logisch identische Fallzahlen über alle Tabellen hinweg identisch bleiben. So nehmen Randsummen einer Kreuztabelle, beispielsweise „Bundesland x Alter", immer die gleichen Werte an, die auch in den entsprechenden Fallzahltabellen der beiden Merkmale ausgegeben werden. Das gilt unabhängig davon, ob der Wert als Innen- oder als Randfeld einer Tabelle auftritt.

Die folgenden Abschnitte erläutern die Funktionsweise der Cell-Key-Methode Schritt für Schritt. Hierbei ist zu beachten, dass ein großer Vorteil der Cell-Key-Methode gerade darin besteht, dass die eigentliche Geheimhaltung von Fallzahltabellen weitgehend automatisiert erfolgt.

3.1 Record Keys

Für die Anwendung der Cell-Key-Methode wird an den Ausgangsdatensatz zunächst ein zusätzliches Merkmal angespielt, das den sogenannten Record Key enthält. Dieser besteht aus einer Zufallszahl zwischen 0 und 1, die jeder Beobachtungseinheit fest zugeordnet wird.

Übersicht 1
Anfügen der Record Keys an den Ausgangsdatensatz

ID	Alter	Einkommen	Record Key
1	jung	mittel	0,54
2	jung	hoch	0,68
3	alt	niedrig	0,14
4	alt	mittel	0,93
5	jung	mittel	0,51
6	alt	mittel	0,37
7	alt	niedrig	0,84
8	alt	hoch	0,19
9	alt	mittel	0,26
10	jung	hoch	0,43
11	alt	mittel	0,99
12	jung	mittel	0,74
13	jung	mittel	0,65
14	alt	niedrig	0,79
15	alt	mittel	0,25

Ausgangsdatensatz

Angespielter Record Key

Zur Veranschaulichung zeigt Übersicht 1 Daten für eine fiktive Gemeinde, in der das Alter und das Einkommen aller erwachsenen Bewohner klassiert erfasst ist. Aus Gründen der Übersichtlichkeit wird ein Record Key mit zwei Nachkommastellen vergeben, in echten Anwendungsfällen ist die Anzahl der Nachkommastellen in der Regel höher.

3.2 Übergangsmatrix

In der Übergangsmatrix wird festgelegt, mit welchem Wert eine Fallzahl überlagert wird (Kleber/Gießing, 2018). Dafür wird für jede Originalfallzahl beschlossen, mit welcher Wahrscheinlichkeit diese zu einem bestimmten anderen Wert verändert wird. In Tabelle 1 bliebe die Originalfallzahl 10 zum Beispiel mit einer Wahrscheinlichkeit von 0,7 eine 10, würde mit einer Wahrscheinlichkeit von je 0,1 zu einer 9 oder 11 und mit einer Wahrscheinlichkeit von je 0,05 zu einer 8 oder 12. So lassen sich verschiedene Rahmenbedingungen festlegen, beispielsweise die maximale Abweichung, die Wahrscheinlichkeit für den Erhalt einer Originalfallzahl, der Ausschluss von 1 und 2 in der überlagerten Tabelle oder eine höhere Bleibewahrscheinlichkeit für höhere Fallzahlen.

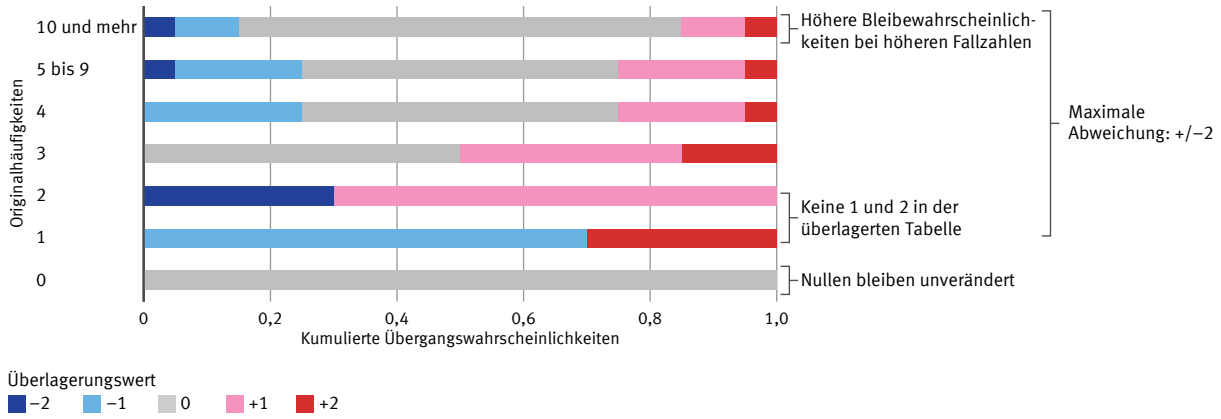
Tabelle 1
Fiktives Beispiel einer Übergangsmatrix

Original-häufigkeit	Zielhäufigkeit																	
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0,7	0	0	0,3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0,3	0	0	0,7	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0,5	0,35	0,15	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0,25	0,5	0,2	0,05	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0,05	0,2	0,5	0,2	0,05	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0,05	0,2	0,5	0,2	0,05	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0,05	0,2	0,5	0,2	0,05	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0,05	0,2	0,5	0,2	0,05	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0,05	0,2	0,5	0,2	0,05	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0,05	0,1	0,7	0,1	0,05	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0,05	0,1	0,7	0,1	0,05	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0,05	0,1	0,7	0,1	0,05	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0,05	0,1	0,7	0,1	0,05	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0,05	0,1	0,7	0,1	0,05	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0,05	0,1	0,7	0,1	0,05

Nullen bleiben unverändert
Keine 1 oder 2 in der überlagerten Tabelle
Maximalabweichung +/−2
Höhere Bleibewahrscheinlichkeit bei höheren Fallzahlen

Dabei ist zu beachten, dass es sich hierbei lediglich um ein fiktives Beispiel einer Übergangsmatrix handelt, um deren mögliche Ausgestaltung vereinfacht darzustellen. Denkbar wäre beispielsweise auch die Festlegung, dass alle ausgewiesenen Nullen echte Nullen sind, dass kein Wert unverändert bleibt oder dass starke Überlagerungen wahrscheinlicher sind als geringe. Da die Übergangsmatrix somit steuert, wie (un-)ähnlich sich die originale und die überlagerte Tabelle sind, stellt sie das Kernstück der Cell-Key-Methode dar, das mit viel Aufwand ausgestaltet wird. Diese tatsächlich verwendeten Übergangsmatrizen unterliegen ebenso wie die festgelegten Rahmenbedingungen der strengen Geheimhaltung und werden an die jeweiligen Bedarfe der Statistiken spezifisch angepasst.

Grafik 1
Fiktives Beispiel eines Überlagerungstableaus



3.3 Überlagerungstableau

Auf Basis der Übergangsmatrix wird ein Überlagerungstableau erstellt (Kleber/Gießing, 2018). Zur Veranschaulichung stellt Grafik 1 das Überlagerungstableau auf Basis der Übergangsmatrix aus Tabelle 1 als Raster dar. Hierfür werden die Wahrscheinlichkeiten der einzelnen Zeilen der Übergangsmatrix kumuliert. Daraus lässt sich im sogenannten Lookup-Schritt ablesen, welcher Überlagerungswert für die jeweilige Originalfallzahl aus der zugehörigen kumulierten Übergangswahrscheinlichkeit resultiert.

Zum besseren Verständnis stellt Übersicht 2 die Veränderungen der Originalfallzahl „4“ exemplarisch dar.

Übersicht 2
Kumulierte Übergangswahrscheinlichkeiten für die Originalfallzahl „4“

Veränderung zu	Entspricht Überlagerung mit	Wahrscheinlichkeit	Kumulierte Wahrscheinlichkeit
3	−1	0,25	0,25
4	0	0,5	0,75
5	1	0,2	0,95
6	2	0,05	1

3.4 Tabellenerstellung

Bei der Erstellung der überlagerten Tabellen kommen schließlich die Originalfallzahl, die kumulierte Übergangswahrscheinlichkeit für diese Originalfallzahl und die eingangs erzeugten Record Keys zusammen, um den Überlagerungswert zu bestimmen:

Im Zuge der Erstellung von Fallzahltabellen mit der Cell-Key-Methode werden nicht nur die Fallzahlen ermittelt. Für jede Tabellenzelle werden darüber hinaus die Record Keys aller Beobachtungseinheiten addiert, die zur entsprechenden Tabellenzelle beitragen. Relevant sind von der Summe der Record Keys allerdings nur die Nachkommastellen, der Wert vor dem Komma wird daher auf 0 gesetzt. So entsteht ein Wert zwischen 0 und kleiner 1, der sogenannte Cell Key.

Übersicht 3 veranschaulicht diesen Mechanismus für die Merkmalskombination „junge Befragte mit mittlerem Einkommen“ sowie für die Summe der Personen mit niedrigem Einkommen.

Die für alle Originalfallzahlen berechneten Cell Keys werden jetzt an das Übergangstableau zurückgespielt. Hier erfolgt der Abgleich, in welchem Raster die kumulierte Wahrscheinlichkeit dem ermittelten Cell Key entspricht. Aus dieser Spalte wird dann der Überlagerungswert für die jeweilige Originalfallzahl abgelesen. Die farbige Markierung des betreffenden Rasters verdeutlicht, mit welchem Wert die Originalfallzahl überlagert wird.

Im Beispiel ergibt sich für die vier jungen Personen mit mittlerem Einkommen aus dem ermittelten Cell Key von 0,44 ein Überlagerungswert von 0. Diese Fallzahl bleibt also unverändert. Für die drei Personen mit niedrigem Einkommen ergibt sich ein Cell Key von 0,77, was laut Überlagerungstableau einer Überlagerung von +1 entspricht. Diese Fallzahl wird also von 3 auf 4 verändert. Siehe Grafik 2

Wird das Verfahren auf alle Felder der Tabelle angewandt, verändert sich die originale Fallzahltabelle anhand des in Übersicht 4 dargestellten Mechanismus.

Der überlagerten Tabelle sieht man zunächst nicht an, dass sie nicht die Originalfallzahlen enthält. Auf den zweiten Blick wird aber schnell deutlich, dass sich die Innenfelder in der Regel nicht zu den Randsummen summieren. Diese und andere Auswirkungen der Cell-Key-Methode stellt in dieser Ausgabe ein weiterer Beitrag vor (Rothe und andere, 2024).

4 Formelle Erläuterung der Cell-Key-Methode

Die Cell-Key-Methode eignet sich in ihrer Grundform für die Geheimhaltung von Fallzahltabellen. Mittlerweile steht auch eine Erweiterung der Cell-Key-Methode auf Wertetabellen zur Verfügung (Gießing/Tent, 2019), die in der deutschen amtlichen Statistik bislang allerdings lediglich für einige Wertmerkmale des Zensus 2022 angewendet wird. Im Folgenden beschränkt sich die formelle Erläuterung der Methodik der Cell-Key-Methode daher auf deren Grundform für Fallzahltabellen in Anlehnung an die Originalveröffentlichung von Fraser/Wooton (2005).

Die Grundidee der Cell-Key-Methode besteht darin, bei den Nutzenden Unsicherheit dahingehend zu erzeugen, ob eine veröffentlichte Fallzahl der Originalfallzahl entspricht oder diese leicht verändert wurde. Dazu erfolgt eine Perturbation aller Originalfallzahlen mittels eines Überlagerungswertes. Bezeichnet man mit $d_i \in \mathbb{Z}$ den dem i-ten Tabellenfeld mit Originalfallzahl $j_i \in \mathbb{N}_0$ zugeordneten Überlagerungswert, so ergibt sich die veränderte Fallzahl k_i gemäß

$$k_i = j_i + d_i .$$

Übersicht 3

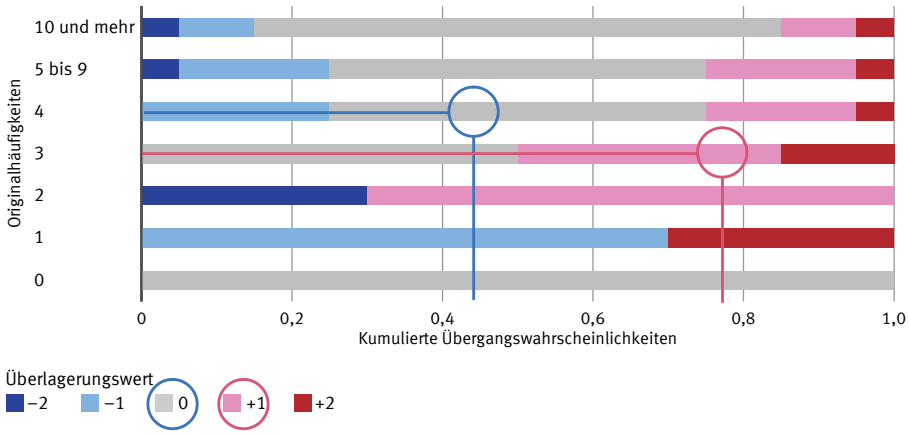
Erstellen der Cell Keys für zwei Beispiele durch die Addition der zugehörigen Record Keys (RK) und das Auf-null-Setzen der Zahl vor dem Komma

ID	Alter	Einkommen	Record Key
1	jung	mittel	0,54
2	jung	hoch	0,68
3	alt	niedrig	0,14
4	alt	mittel	0,93
5	jung	mittel	0,51
6	alt	mittel	0,37
7	alt	niedrig	0,84
8	alt	hoch	0,19
9	alt	mittel	0,26
10	jung	hoch	0,43
11	alt	mittel	0,99
12	jung	mittel	0,74
13	jung	mittel	0,65
14	alt	niedrig	0,79
15	alt	mittel	0,25

		Alter		
		jung	alt	Summe
Einkommen	niedrig	0	3	3 ∑RK = 0,14+0,84+ 0,79 = 1,77 ➔Cell Key = 0,77
	mittel	4 ∑RK = 0,54+0,51+ 0,74+0,65 = 2,44 ➔Cell Key = 0,44	5	9
	hoch	2	1	3
	Summe	6	9	15

Grafik 2

Ablesen ("Lookup") der Überlagerungswerte aus dem Überlagerungstableau für die 4 jungen Personen mit mittlerem Einkommen und einem Cell Key von 0,44 (blau) und den 3 Personen mit niedrigem Einkommen und einem Cell Key von 0,77 (rot)



Übersicht 4

Darstellung der Arbeitsschritte der Cell-Key-Methode

Ausgangstabelle

Originale Tabelle

		Alter		
		jung	alt	Σ
Einkommen	niedrig	0	3	3
	mittel	4	5	9
	hoch	2	1	3
	Σ	6	9	15

Addition der Record Keys, nur Nachkommastellen

Interne Vorgänge

Cell Keys

		Alter		
		jung	alt	Σ
Einkommen	niedrig	0,00	0,77	0,77
	mittel	0,44	0,80	0,24
	hoch	0,11	0,19	0,30
	Σ	0,55	0,76	0,31

Ablesen aus Überlagerungstableau

Überlagerungswerte

		Alter		
		jung	alt	Σ
Einkommen	niedrig	0	+1	+1
	mittel	0	+1	−1
	hoch	−2	−1	0
	Σ	0	+1	0

Addition der Überlagerungswerte

Finale Tabelle

Überlagerte Tabelle

		Alter		
		jung	alt	Σ
Einkommen	niedrig	0	4	4
	mittel	4	6	8
	hoch	0	0	3
	Σ	6	10	15

Wie bei allen datenverändernden Verfahren gehört zu den wesentlichen Eigenschaften der Cell-Key-Methode, dass alle Originalfallzahlen unabhängig ihrer Kritikalität beziehungsweise ihres Aufdeckungsrisikos verändert werden können.

4.1 Anforderungen an die Überlagerungswerte

Für die Überlagerungswerte d_i gelten für alle i folgende Eigenschaften:

a) **Unverzerrtheit:** $E[d_i] = 0$, das heißt aus der Überlagerung der Tabellenfelder ergibt sich keine systematische Verzerrung des Gesamtergebnisses.

b) **Nicht-Negativität:** $d_i \geq -j_i$, das heißt für jedes Tabellenfeld wird sichergestellt, dass sich durch die Überlagerung keine negative Fallzahl ergibt ($\forall i: k_i \geq 0$).

c) **Ganzzahligkeit:** $d_i \in \mathbb{Z}$, das heißt die Überlagerungswerte müssen ganzzahlig sein, damit auch für die veränderten Fallzahlen Ganzzahligkeit gewährleistet ist. Aus den Bedingungen (b) und (c) ergibt sich somit $k_i \in \mathbb{N}_0$.

4.2 Parameter der Cell-Key-Methode

Die Anwendung der Cell-Key-Methode setzt die Festlegung von Parametern und Bedingungen voraus, durch welche das für die Zuweisung der Überlagerungswerte maßgebende stochastische Modell determiniert wird. Dabei wird zwischen zwingend festzulegenden Parametern sowie zusätzlichen optional zu setzenden Restriktionen unterschieden. Die konkrete Ausgestaltung der Parametrisierung erfolgt dabei stets durch die für die betreffende Statistik zuständige Fachseite.

Konkret ist die Festlegung zweier Parameter obligatorisch:

- **Maximale Abweichung** $D \in \mathbb{N}$ zwischen originalen und veränderten Fallzahlen, sodass $\forall i: |d_i| \leq D$.
- **Varianz** V der Abweichungen von den Originalfallzahlen: Die Varianz entspricht

$$V := Var[d_i] = E[d_i^2] = \sum_{i=-D}^{D} (p_i \cdot d_i^2).$$

Durch die Festlegung von D ergibt sich als Wertebereich für V implizit $(0; D^2]$.

Die Festlegung der maximalen Abweichung zwischen originalen und veränderten Fallzahlen dient insbesondere der Steuerung des durch die Datenveränderung induzierten Informationsverlustes und damit der Qualität der geheim gehaltenen Ergebnisse. Mit der Varianz der Abweichungen von den Originalfallzahlen wird die Unsicherheit kalibriert, die bei den Nutzenden durch die Datenveränderung erzeugt werden soll. Eine geringe Varianz impliziert dabei, dass ein hoher Anteil der Originalfallzahlen nur geringfügig oder gar nicht (also $d_i = 0$) verändert wird. Je größer V gewählt wird, umso größer ist die Wahrscheinlichkeit, dass (unter Berücksichtigung der festgelegten Maximalabweichung D) hohe Abweichungen zwischen originalen und veränderten Fallzahlen auftreten.

4.3 Übergangsmatrix

Die maximale Abweichung von der Originalfallzahl sowie die Varianz der Überlagerungswerte bestimmen maßgeblich die Wahrscheinlichkeiten, mit der ein bestimmter Überlagerungswert einer Originalfallzahl zugeordnet wird. Im Folgenden bezeichnet $p_{jk} \in [0;1]$ die Wahrscheinlichkeit, dass die Originalfallzahl j zur Fallzahl k verändert wird (beziehungsweise den Überlagerungswert $d = |j-k|$ erhält). Da diese Wahrscheinlichkeit nur von der Höhe der Originalfallzahl abhängt, jedoch nicht von einem konkreten Tabellenfeld i, kann hier auf den Index i verzichtet werden. Die (bedingten) Übergangswahrscheinlichkeiten werden für alle Kombinationen möglicher Originalfallzahlen und veränderter Fallzahlen in der sogenannten Übergangsmatrix **T** gesammelt:

$$\mathbf{T} = \begin{pmatrix} p_{00} & p_{01} & p_{02} & \cdots & p_{0k} & \cdots \\ p_{10} & p_{11} & p_{12} & \cdots & p_{1k} & \cdots \\ p_{20} & p_{21} & p_{22} & \cdots & p_{2k} & \cdots \\ \vdots & \vdots & \vdots & \ddots & & \\ p_{j0} & p_{j1} & p_{j2} & & p_{jk} & \\ \vdots & \vdots & \vdots & & & \ddots \end{pmatrix}$$

T ist eine quadratische Matrix, die folgende Eigenschaften aufweist:

- **T** enthält auf der Hauptdiagonalen die sogenannten Bleibewahrscheinlichkeiten p_{jj} mit $j = k$, das heißt die Wahrscheinlichkeiten, dass eine Originalfallzahl j unverändert bleibt.

- Zeilenweise enthält **T** die bedingten Wahrscheinlichkeitenverteilungen für die Überlagerungen der Originalfallzahlen j, das heißt $\forall j: \sum_k p_{jk} = 1$.

- Durch die zwingende Festlegung einer maximalen Abweichung D von den jeweiligen Originalfallzahlen ist $p_{jk} = 0$ für alle $k \notin \{j - D, \ldots, j + D\}$.

- Für kleine Fallzahlen $j < D$ kann das eigentlich durch D festgelegte Intervall für die möglichen Fallzahlen nach Datenveränderung $[j - D, \ldots, j + D]$ nicht vollständig ausgeschöpft werden, um der Bedingung der Nicht-Negativität der veränderten Fallzahlen zu genügen. Das tatsächliche Intervall für die möglichen Fallzahlen nach Veränderung der Originalfallzahl j lautet somit $\Pi = [\max\{j - D; 0\}, \ldots, j + D]$. Negative Überlagerungswerte sind bei kleinen Originalfallzahlen (bei gleichzeitiger Gewährleistung der Unverzerrtheit der veränderten Fallzahlen) somit nur eingeschränkt möglich (Höhne / Höninger, 2018).

Aaron / stock.adobe.com

4.4 Weitere optionale Restriktionen zur Kalibrierung der Übergangsmatrix

Neben den oben genannten zwingend festzulegenden Parametern D und V können optional weitere Nebenbedingungen formuliert werden, welche die Gestalt der Übergangsmatrix beeinflussen. Mögliche Restriktionen lauten:

- **Festlegung einer Bleibewahrscheinlichkeit:** Es kann festgelegt werden, dass ein bestimmter Anteil P_V aller Originalfallzahlen unverändert bleiben soll. In diesem Fall entsprechen $P_V \cdot 100\,\%$ aller Fallzahlen in den geheim gehaltenen Tabellen ihrem jeweiligen Originalwert.

- **Original-Nullen sollen unverändert bleiben:** $p_{00} = 1$ beziehungsweise $\forall k > 0: p_{0k} = 0$. In der Realität nicht existierende Ausprägungen von Merkmalskombinationen bleiben auch nach Anwendung der Cell-Key-Methode ausgeschlossen.

- **Ausschluss kleiner Fallzahlen nach Datenveränderung:** Die Ausgabe sehr kleiner veränderter Fallzahlen kann optional bis einschließlich eines Schwellenwerts $m > 0$ ausgeschlossen werden. Für die entsprechenden veränderten Fallzahlen $k = 1, \ldots, m$ gilt dann $\forall j: p_{jk} = 0$. Die entsprechenden Spaltenvektoren für die ausgeschlossenen veränderten Fallzahlen in $\mathbf{T}$ entsprechen dann dem Nullvektor. Häufig wird fachseitig die 1 als veränderte Fallzahl ausgeschlossen ($\forall j: p_{j1} = 0$).

- **Gewährleistung einer symmetrischen Verteilung:** $\forall d \in \{-D, \ldots, D\}: p_{j(j-d)} = p_{j(j+d)}$, das heißt Veränderungen einer Originalfallzahl um die Überlagerungswerte $-d$ und d besitzen die identische Wahrscheinlichkeit. Hieraus ergibt sich eine symmetrische Verteilung der Übergangswahrscheinlichkeiten für jede Fallzahl j um die entsprechende Bleibewahrscheinlichkeit p_{jj}. Dabei ist zu beachten, dass die Symmetrieeigenschaft der Überlagerungsverteilung nur für solche Originalfallzahlen j gewährleistet werden kann, für welche $j \geq D$ gilt. Werden zusätzlich kleine veränderte Fallzahlen bis einschließlich des Schwellenwerts m ausgeschlossen, erweitert sich diese Bedingung zu $j > D + m$.

4.5 Berechnung der Übergangsmatrix

Auf Basis der obligatorisch festzulegenden Parameter sowie unter Berücksichtigung der weiteren optionalen Restriktionen an die Ausgestaltung der Übergangsmatrix ist für jede Originalfallzahl die konkrete (bedingte) Verteilung der Überlagerungswerte zu berechnen. Marley/Leaver (2011) schlagen hierzu die Maximierung der Entropie der (bedingten) Überlagerungsverteilungen vor. Die Entropie stellt ein Streuungsmaß einer Wahrscheinlichkeitsverteilung dar, dessen Maximierung in diesem Zusammenhang als Minimierung des Aufdeckungsrisikos interpretiert werden kann (Marley/Leaver, 2011). Bezeichnet Π die Menge der für die betreffende Originalfallzahl möglichen Fallzahlen nach Datenveränderung, so lautet das Maximierungsproblem zur Berechnung der Überlagerungsverteilung für Originalfallzahl j

$$\max_{\mathbf{p}_{jk}} \left\{ -\sum_{k \in \Pi} \left(p_{jk} \cdot \log_2 p_{jk} \right) \right\}.$$

Einschließlich aller formulierten Nebenbedingungen ergibt sich somit ein nicht-lineares Gleichungssystem. Gießing (2016) stellt einen Ansatz zur Lösung des Optimierungsproblems mittels eines Lagrange-Ansatzes vor, welcher im R-Paket *ptable* (Enderle, 2023) zur Erstellung von CKM-Übergangsmatrizen angewendet wird. Weitere Ausführungen zur Maximierung der Entropie sind Enderle/Vollmar (2019) zu entnehmen.

Für alle Originalfallzahlen $j > m + D$ ist die Überlagerungsverteilung strukturell identisch, das heißt $\forall l \geq 0: p_{jk} = p_{(j+l)(k+l)}$. Die resultierende Überlagerungsverteilung für $j = m + D + 1$ gilt somit – jeweils um $a \in \mathbb{N}$ Spalten in **T** nach rechts verschoben – auch für alle größeren Originalfallzahlen $j + a$.

4.6 Record Keys und Cell Keys

Nach der erfolgten Spezifizierung der Übergangsmatrix **T** werden den Originalfallzahlen konkrete Überlagerungswerte zugeordnet. Dies erfolgt anhand des vorliegenden Datenmaterials.

Dazu wird zunächst jedem Merkmalsträger $s = 1, \ldots, S$ im Mikrodatensatz eine feste Zufallszahl $r_s \sim U[0; 1]$ zugewiesen. Die Zufallszahlen werden aus einer stetigen Gleichverteilung gezogen und als Record Keys bezeichnet. Der einem Merkmalsträger zugewiesene Record Key bleibt für alle Auswertungen, die für die betroffene Statistik erstellt werden, mindestens für die laufende Berichtsperiode identisch.

Auf Ebene der Tabellenfelder wird anhand der Record Keys eine „Kennziffer" für jedes einzelne Tabellenfeld gebildet, welche zur Zuweisung des Überlagerungswertes für das betreffende Tabellenfeld verwendet wird. Dieser sogenannte Cell Key c_i wird für Tabellenfeld i gemäß

$$c_i = \sum_{s \in I} r_s - \left\lfloor \sum_{s \in I} r_s \right\rfloor \sim U[0; 1)$$

berechnet, wobei die Menge I alle Merkmalsträger s enthält, die zu Tabellenfeld i beitragen. Zur Berechnung des Cell Keys für Tabellenfeld i wird die Summe der Record Keys aller zu i beitragenden Merkmalsträger um deren nächst kleineren ganzzahligen Betrag reduziert (hinterer Term mit unterer Gauß-Klammer). Diese Rechenoperation ist notwendig, da gleichverteilte Zufallsvariablen ihre Verteilungseigenschaft bei Summierung (vorderer Term) verlieren und durch die Korrektur die Eigenschaft einer stetigen Gleichverteilung für die Cell Keys wiederhergestellt wird. Im Ergebnis weist nach diesem Schritt jedes zu überlagernde Tabellenfeld einen spezifischen Cell Key mit einem Wert $c_i \in [0; 1)$ auf, der von den konkreten Merkmalsträgern beziehungsweise deren Record Keys abhängt, die zum Tabellenfeld beitragen. Durch die hier dargestellte Vorgehensweise wird tabellenübergreifende Konsistenz der veränderten Fallzahlen sichergestellt, da logisch identische Tabellenfelder (das heißt mit identischen beitragenden Merkmalsträgern) stets den gleichen Cell Key erhalten.

4.7 Zuweisung der Überlagerungswerte

In einem letzten Schritt werden die generierten Cell Keys genutzt, um anhand der spezifizierten Übergangsmatrix zu entscheiden, welcher konkrete Überlagerungswert einem Tabellenfeld zugewiesen wird.

Die Übergangsmatrix wird dazu in ein sogenanntes Überlagerungstableau überführt, indem die Überlagerungsverteilung für jede Originalfallzahl j schrittweise aggregiert wird. Dazu sei die Verteilungsfunktion F_j gemäß

$$F_j(d) := \sum_{b \leq d;\, d \in \Lambda} p_{b|j}$$

definiert, wobei die Menge $\Lambda_j = \{d_1, d_2, \ldots, d_{n-1}, d_n\}$ alle für j infrage kommenden Überlagerungswerte enthalte und $p_{d|j}$ die Wahrscheinlichkeit einer Überlagerung

der Originalfallzahl j mit Überlagerungswert d bezeichnet. Die Höhe der einzelnen Übergangswahrscheinlichkeiten wird dabei über die Breite der Intervalle $[0;F_j(d_1)]$, $(F_j(d_1); F_j(d_2)]$, …, $(F_j(d_{n-1}); F_j(d_n)]$ abgebildet, wobei $F_j(d_n)] = 1$ ist. Die Vereinigungsmenge aller Intervalle deckt das Intervall [0; 1] somit vollständig und überlappungsfrei ab.

Da die Cell Keys auf dem Intervall [0; 1) gleichverteilt sind, kann die Zuweisung des Überlagerungswerts auf Basis eines einfachen Abgleichs zwischen dem Cell Key eines Tabellenfeldes und der Verteilungsfunktion der Überlagerungswerte vorgenommen werden. Dabei wird der Überlagerungswert d_i zum Tabellenfeld i mit Cell Key c_i anhand des folgenden Mechanismus zugeordnet:

$$d_i = d_r | d_r \in \Lambda : c_i \in (F_j(d_{r-1}); F_j(d_r)]$$

Als Überlagerungswert für das Tabellenfeld i wird somit der Überlagerungswert d_r ausgewählt, falls der Cell Key des Tabellenfeldes in das Teilintervall fällt, welches durch die Verteilungsfunktionswerte von d_{r-1} und d_r aufgespannt wird (und dessen Breite der Wahrscheinlichkeit einer Überlagerung mit d_r entspricht). Dieser Zuweisungsmechanismus wird auf alle zu überlagernden Tabellenfelder angewendet. Durch die Gleichverteilungseigenschaft der Cell Keys ist gewährleistet, dass über alle Tabellen hinweg der Anteil an mit einem bestimmten Überlagerungswert überlagerten Tabellenfeldern der jeweiligen Übergangswahrscheinlichkeit $p_{d|j} \cdot 100\,\%$ entspricht.

5. Fazit

Mit der Cell-Key-Methode hält ein neues Geheimhaltungsverfahren Einzug in die amtliche Statistik und damit auch in die Forschungsdatenzentren. Die Cell-Key-Methode ist ein Verfahren, das auf einer posttabularen stochastischen Überlagerung basiert. Die feste Zuordnung eines Record Keys zu jeder Beobachtungseinheit stellt sicher, dass Veränderungen der originalen Fallzahlen konsistent und über verschiedene Ergebnisläufe hinweg replizierbar erfolgen. Dies geht jedoch zulasten der Additivität der Ergebnistabellen. Welche Auswirkungen die Anwendung der Cell-Key-Methode auf Tabellenergebnisse und darauf basierende Kennzahlen darüber hinaus hat, beschreibt im Detail der Artikel „Die Cell-Key-Methode in den Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder – Teil 2: Auswirkungen des neuen Geheimhaltungsverfahrens“ in Ausgabe 9/2024 dieser Zeitschrift (Rothe und andere, 2024). ■

Literatur

Enderle, Tobias/Vollmar, Meike. Geheimhaltung in der Hochschulstatistik. In: WISTA Wirtschaft und Statistik. Ausgabe 6/2019, Seite 87 ff. Verfügbar unter: www.destatis.de/DE/Methoden/WISTA-Wirtschaft-und-Statistik/2019/06/geheimhaltung-hochschulstatistik-062019.pdf?__blob=publicationFile

Enderle, Tobias. ptable: Generation of Perturbation Tables for the Cell-Key Method. R package version 1.0.0. 2023. [Zugriff am 30. April 2024]. Verfügbar unter: https://cran.r-project.org/web/packages/ptable/index.html

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 30. April 2024]. Verfügbar unter: https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2005/wp.35.e.pdf

Giessing, Sarah/Tent, Reinhard. Concepts for generalising tools implementing the cell key method to the case of continuous variables. In: Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality. Den Haag 2019. [Zugriff am 30. April 2024]. Verfügbar unter: https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2019/mtg1/SDC2019_S2_Germany_Giessing_Tent_AD.pdf

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. DOI: 10.1007/978-3-319-45381-1_18

Höhne, Jörg. Das Geheimhaltungsverfahren SAFE. In: Zeitschrift für amtliche Statistik Berlin Brandenburg. Ausgabe 2/2015, Seite 16 ff. [Zugriff am 30. April 2024]. Verfügbar unter: www.statistischebibliothek.de/mir/receive/BBHeft_mods_00016121

Höhne, Jörg/Höninger, Julia. Die Cell-Key-Methode – ein Geheimhaltungsverfahren. In: Zeitschrift für amtliche Statistik Berlin Brandenburg. Ausgabe 3+4/2018, Seite 14 ff. [Zugriff am 30. April 2024]. Verfügbar unter: www.statistischebibliothek.de/mir/receive/BBHeft_mods_00036268

Kleber, Birgit/Gießing, Sarah. Geheimhaltung beim Zensus 2021. In: Methoden – Verfahren – Entwicklungen. Nachrichten aus dem Statistischen Bundesamt. Ausgabe 2/2018, Seite 3 ff. [Zugriff am 30. April 2024]. Verfügbar unter: www.destatis.de/DE/Methoden/Qualitaet/methoden-verfahren-entwicklung-02_2018.pdf?__blob=publicationFile&v=1

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. [Zugriff am 30. April 2024]. Verfügbar unter: https://2011.isiproceedings.org/papers/450007.pdf

Rohde, Johannes/Seifert, Christiane/Gießing, Sarah/Setzer, Stefanie (unter Mitarbeit von Breitenfeld, Jörg/Brings, Stefan/Höhne, Jörg/Höninger, Julia/Rothe, Patrick/Schedding-Kleis, Ulrike). Entscheidungskriterien für die Auswahl eines Geheimhaltungsverfahrens. Version 1.1 vom 23.04.2021. Internes Dokument des Statistischen Verbunds (Statistische Ämter des Bundes und der Länder).

Rothe, Patrick/Güttgemanns, Volker/Rohde, Johannes/Setzer, Stefanie. Die Cell-Key-Methode in den Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder. – Teil 2: Auswirkungen des neuen Geheimhaltungsverfahrens. In: WISTA Wirtschaft und Statistik. Ausgabe 3/2024, Seite 45 ff. Verfügbar unter: www.destatis.de/DE/Methoden/WISTA-Wirtschaft-und-Statistik/2024/03/cell-key-methode-teil2-032024.pdf?__blob=publicationFile

Rechtsgrundlagen

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 14 des Gesetzes vom 8. Mai 2024 (BGBl. I Nr. 152) geändert worden ist.