Rothe, Patrick; Güttgemanns, Volker; Rohde, Johannes; Setzer, Stefanie

Article

Die Cell-Key-Methode in den Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder – Teil 2: Auswirkungen des neuen Geheimhaltungsverfahrens

WISTA - Wirtschaft und Statistik

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

Suggested Citation: Rothe, Patrick; Güttgemanns, Volker; Rohde, Johannes; Setzer, Stefanie (2024) : Die Cell-Key-Methode in den Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder – Teil 2: Auswirkungen des neuen Geheimhaltungsverfahrens, WISTA - Wirtschaft und Statistik, ISSN 1619-2907, Statistisches Bundesamt (Destatis), Wiesbaden, Vol. 76, Iss. 3, pp. 45-54

This Version is available at:
https://hdl.handle.net/10419/299143

DIE CELL-KEY-METHODE IN DEN FORSCHUNGSDATENZENTREN DER STATISTISCHEN ÄMTER DES BUNDES UND DER LÄNDER

Teil 2: Auswirkungen des neuen Geheimhaltungsverfahrens

Patrick Rothe, Volker Güttgemanns, Johannes Rohde, Stefanie Setzer

↘ **Schlüsselwörter:** Geheimhaltung – stochastische Überlagerung – post-tabular – Verhältniszahlen – Zeitreihen

ZUSAMMENFASSUNG

Die Cell-Key-Methode ist ein hauptsächlich für die Geheimhaltung von Fallzahltabellen entwickeltes Geheimhaltungsverfahren. In den Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder werden häufig umfangreiche Analysen vorgenommen, auf deren Ergebnisdarstellung sich das neue Verfahren ebenfalls auswirken kann. Der Artikel beschreibt die Auswirkungen, die das Verfahren auf die Ergebnisqualität, Fallzahltabellen, Verhältniszahlen und Zeitreihen hat.

↘ ***Keywords:*** *confidentiality – stochastic perturbation – post-tabular – ratios – time series*

ABSTRACT

The cell key method is a disclosure control method that was primarily developed to ensure the confidentiality of frequency tables. Extensive analyses are frequently carried out in the Research Data Centres of the statistical offices of the Federation and the Länder, and the new method can affect the presentation of the results. This article describes the effects that the method has on the quality of results, frequency tables, ratios and time series.

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.

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.

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“.

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.

1

Einleitung

Die Statistischen Ämter des Bundes und der Länder führen für ausgewählte Statistiken ein neues Geheimhaltungsverfahren ein: die Cell-Key-Methode (CKM)[1]. Um die Geheimhaltung der Ergebnisse über alle Veröffentlichungen hinweg sicherzustellen, wenden die Forschungsdatenzentren der amtlichen Statistik dieses Verfahren entsprechend an. Der Artikel „Die Cell-Key-Methode in den Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder – Teil 1: Vorstellung des neuen Geheimhaltungsverfahrens“ (Setzer und andere, 2024) beschreibt das Verfahren ausführlich.

Die Schutzwirkung der Cell-Key-Methode entsteht durch eine post-tabulare Überlagerung aller Fallzahlen. In der überlagerten Tabelle lässt sich hierdurch letztendlich nicht mehr erkennen, welche Werte überlagert wurden und welche weiterhin dem Originalwert entsprechen. Diese Vorgehensweise stellt einen großen Unterschied zur bisher überwiegend angewandten Zellsperrung dar, in der ausgewiesene Werte immer dem Originalwert entsprechen, insbesondere kleine Fallzahlen aber gesperrt werden. Die Auswirkungen des neuen Verfahrens auf die Ergebnisqualität, Fallzahltabellen, Verhältniszahlen und Zeitreihen werden in diesem Beitrag vorgestellt. Dazu beschreibt Kapitel 2 detailliert die Auswirkungen auf Fallzahltabellen, Kapitel 3 befasst sich mit den Effekten, die die Cell-Key-Methode auf Verhältniszahlen haben kann. Auch auf Zeitreihen kann sich die Anwendung der Cell-Key-Methode auswirken, wie in Kapitel 4 dargestellt wird. Ein kurzes Fazit zur Nutzung des neuen Geheimhaltungsverfahrens in den Forschungsdatenzentren beschließt den Artikel.

1 Die Cell-Key-Methode wurde vom australischen Statistikamt (Australien Bureau of Statistics) entwickelt (Fraser/Wooton, 2005).

2

Auswirkungen auf Fallzahltabellen

2.1 Kurzüberblick

Bei Fallzahltabellen, die mit der Cell-Key-Methode überlagert wurden, sind zwei wesentliche Auswirkungen besonders hervorzuheben:

- Mit der Cell-Key-Methode geheim gehaltene Tabellen sind immer konsistent. Kommen logisch identische Tabellenfelder also in verschiedenen Tabellen vor (zum Beispiel in einer univariaten Fallzahltabelle und als Randwert einer Kreuztabelle), wird immer die gleiche überlagerte Fallzahl ausgegeben. Dies gewährleisten eine vorher festgelegte Übergangsmatrix und der Record Key, der den einzelnen Erhebungseinheiten fest zugeordnet ist.
- Mit der Cell-Key-Methode geheim gehaltene Tabellen sind nicht additiv. Da Tabelleninnen- und -randfelder unabhängig voneinander überlagert werden, addieren sich die überlagerten Innenfelder nicht (oder nur zufällig) zu den überlagerten Randsummen. Eine Wiederherstellung der Additivität wäre zwar theoretisch möglich, würde aber die Konsistenz und Qualität der Ergebnisse beeinträchtigen und zudem die benötigte Rechenzeit erhöhen, sodass darauf verzichtet wird.

2.2 Auswirkungen im Detail

Additivität und Konsistenz der geheim gehaltenen Tabellenergebnisse stellen zwei wesentliche Anforderungen an statistische Verfahren zur Geheimhaltung von Tabellen dar. Gerade im Kontext der Verwendung der Cell-Key-Methode sind diese beiden Eigenschaften von besonderer Bedeutung.

Additivität einer Tabelle ist dann gegeben, wenn sich die Innenfelder der Tabelle zur ausgewiesenen Summe im entsprechenden Randfeld aufaddieren lassen – was bei einer unbearbeiteten Tabelle der Normalfall ist und in aller Regel von den Nutzenden auch so erwartet wird. Bestimmte Verfahren zur statistischen Geheimhaltung, die aus der Familie der datenverändernden Methoden stammen, führen im Ergebnis jedoch dazu, dass diese

Eigenschaft verletzt wird. Somit entspricht die Summe der aufaddierten Innenfelder nicht zwangsläufig der in der Tabelle ausgewiesenen Randsumme. Dieser geschilderte Effekt („Nicht-Additivität“) tritt auch bei der Cell-Key-Methode auf und entsteht, indem jedes Feld einer Tabelle separat – das heißt unabhängig von allen anderen Tabellenfeldern – dem datenverändernden Algorithmus unterzogen wird. Innenfelder werden somit genauso wie die in der Tabelle enthaltenen Zwischen- oder Gesamtsummen behandelt. Dieses separate Vorgehen bewirkt in der Regel einen Verlust der Additivitätseigenschaft.

Auf den ersten Blick kann dieser Effekt für die Nutzenden irritierend erscheinen, letztlich führt dieses Vorgehen unter Gesichtspunkten der Datenqualität und Informationserhaltung jedoch zu einem besseren Ergebnis: Die separate Überlagerung jedes Tabellenfeldes verhindert, dass sich Abweichungen über eine Tabellenzeile oder -spalte hinweg aufaddieren und die nach Geheimhaltung ausgewiesene Randsumme gegenüber dem Originalwert stark abweicht. Eine Veränderung um die Höhe der Maximalabweichung multipliziert mit der Anzahl der beteiligten Tabellenfelder wäre dabei im Extremfall für Randsummen nicht ausgeschlossen. Die unabhängige Überlagerung aller Tabellenfelder führt jedoch auch bei Tabellenfeldern mit Zwischen- oder Gesamtsummen dazu, dass der veränderte Wert vom Originalwert niemals weiter abweichen kann als es die vorher festgelegte Maximalvorgabe zulässt.[2]

Dieser Vorteil hinsichtlich der Datenqualität wird jedoch durch die Diskrepanz zwischen den aufaddierten Summen der einzelnen Tabellenfelder und den in der Tabelle ausgewiesenen Zwischen- und Gesamtsummen erkauft. Dieser Unterschied kann unter Umständen deutlich ausfallen. Daher sollten Nutzende darauf verzichten, selbstständig Rechenoperationen mit den Angaben aus den per Cell-Key-Methode geheim gehaltenen Ergebnistabellen vorzunehmen, sondern direkt auf die in der Tabelle ausgewiesenen Summenfelder zurückgreifen.

Eine Nicht-Additivität fällt immer dann besonders auf, wenn nur sehr wenige Tabellenfelder zu einer Randsumme beitragen und diese durch einfaches Kopfrechnen sehr schnell ermittelt werden kann. Ein Beispiel hierfür ist die Aufgliederung des Merkmals „Geschlecht“ nach den drei Merkmalsausprägungen „weiblich“, „männlich“ und „divers“ in einer fiktiven Tabelle: Hierbei ist die Summation der Fallzahlen in den drei Innenfeldern sehr einfach möglich, wobei es sehr wahrscheinlich ist, dass die selbstständig berechnete Summe von der in der geheim gehaltenen Tabelle ausgewiesenen Gesamtsumme abweicht. In größeren Tabellen mit einer größeren Anzahl an beitragenden Spalten oder Zeilen fällt dieses Problem jedoch nicht direkt ins Auge.

Konsistenz hingegen bezeichnet die Eigenschaft eines spezifischen Tabellenfeldes, immer den identischen inhaltlichen Wert auszuweisen – unabhängig von der konkreten Tabelle, in der diese Merkmalskombination ausgewiesen wird. Nicht alle Geheimhaltungsverfahren können dies gewährleisten, da hierfür zwingend gesichert sein muss, dass ein- und dasselbe inhaltliche Tabellenfeld in allen Fällen durch die gewählte Methode identisch behandelt wird. Die Cell-Key-Methode ist aufgrund ihrer Ausgestaltung in der Lage, diese Eigenschaft zu erfüllen. Gewährleistet wird dies durch die Aggregation der zu einem individuellen Tabellenfeld beitragenden Record Keys zum für das Verfahren namensgebenden Cell Key. Da dieser bei Vorhandensein derselben Kombination von Merkmalsträgern jeweils identisch ausfällt, ist auch die Datenveränderung stets identisch. Das gilt unabhängig davon, in welcher Tabelle und auf welchem Veröffentlichungsweg das entsprechende Tabellenfeld publiziert wird.

Neben einer höheren Nutzendenfreundlichkeit – die Ergebnisse hängen beispielsweise nicht vom Abrufzeitpunkt ab – geht diese Eigenschaft auch mit einem gesteigerten Schutz der Daten und der für deren Überlagerung genutzten Parameter einher. Bei nicht konsistentem Verhalten könnte im Gegensatz hierzu bei einer größeren Anzahl an Ergebnisabrufen – zumindest näherungsweise – auf die dahinterliegende Originalangabe geschlossen werden, da sich die positiven und negativen Abweichungen vom sich ergebenden Mittelwert bei einem erwartungstreuen Verfahren im Mittel ausgleichen.

2 Weitere Betrachtungen zur Nicht-Additivität der Cell-Key-Methode finden sich in Höhne/Höninger (2018).

Übersicht 1
Auswirkungen der Cell-Key-Methode auf die Additivität einer Fallzahltabelle

Originaltabelle

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

Cell-Key-Methode →

Überlagerte Tabelle

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

Farblegende

		Randsumme ist Summe der Innenfelder	
		ja	nein
Ausgangswert unverändert	ja		
	nein		

↘ Übersicht 1 basiert auf dem Beispiel des Artikels „Die Cell-Key-Methode in den Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder – Teil 1: Vorstellung des neuen Geheimhaltungsverfahrens“ (Setzer und andere, 2024) und zeigt die Veränderungen, die durch die Cell-Key-Methode bezüglich der Randsummen ausgelöst werden können.

2.3 Weitere Qualitätskriterien

Ein Geheimhaltungsverfahren sollte stets so konzipiert sein, dass das Gleichgewicht zwischen dem Schutz der Angaben der Befragten und dem Erhalt des Informationspotenzials einer Statistik sichergestellt ist. Höhne/Höninger (2018) benennen dafür neben den bereits erläuterten Aspekten der Konsistenz und Additivität drei weitere grundsätzliche Ziele von Geheimhaltungsverfahren, wobei unterschiedliche Verfahren die einzelnen Ziele unterschiedlich priorisieren. ↘ Übersicht 2 stellt diese Ziele vor und geht auf deren Erfüllung durch die Cell-Key-Methode und die aktuell in den Forschungsdatenzentren überwiegend genutzte Zellsperrung ein.

Zur Frage der Akzeptanz der Ergebnisse (Punkt 3): Insgesamt haben interne Testrechnungen der statistischen Ämter ergeben, dass der durch Anwendung der Cell-Key-Methode verursachte Informationsverlust bei der Wahl geeigneter Parameter gering ist (Rohde und andere, 2021). Dabei ist zu beachten, dass sich bei kleinen (Original-)Fallzahlen auch Überlagerungen mit einem absolut gesehen geringen Wert stark auswirken können: Wird beispielsweise die Originalfallzahl 3 mit dem Wert +6 überlagert (+200 %), ist die relative Abweichung deutlich höher als bei der entsprechenden Überlagerung einer höheren Fallzahl (zum Beispiel +0,6 % bei einer Originalfallzahl von 1 000). Die Generierung von Tabellen mit vielen kleinen Fallzahlen ist daher nach Möglichkeit zu vermeiden, indem Auswertungen nicht auf einer unnötig tiefen regionalen oder fachlichen Gliederungsebene vorgenommen werden.

Übersicht 2

Vergleich von Cell-Key-Methode und Zellsperrung bei Erreichen der Geheimhaltungsziele

Ziel	Cell-Key-Methode	Zellsperrung
1. Keine unplausiblen Werte	Durch die Ausgestaltung der Übergangsmatrix kann sichergestellt werden, dass in Fallzahltabellen keine unplausiblen Ergebnisse generiert werden (Originalfallzahl 0 wird nicht verändert). Bei der Berechnung von Kennzahlen kann es jedoch vor allem bei kleinen Fallzahlen zu Unplausibilitäten kommen, wenn beispielsweise Zähler und Nenner von Verhältniszahlen gegenläufig verändert werden oder sich bei Trendanalysen das Vorzeichen ändert. Diese ungewünschten Effekte lassen sich jedoch durch geeignete Maßnahmen vermeiden.	Fallzahlen werden nicht verändert, sondern bei zu geringer Besetzung gesperrt. Daher können keine unplausiblen Fallzahlen entstehen. Dies gilt auch für die Berechnung von Kennzahlen und Zeitreihen.
2. Schutz der Einzelangaben	Bei der Ausgestaltung der Übergangsmatrix wird sichergestellt, dass die Veränderungen groß genug sind, um den Schutz der Einzelangaben sicherzustellen.	Fallzahlen unterhalb einer festgelegten Mindestfallzahl werden gesperrt, durch Gegensperrungen wird eine Rückrechnung verhindert. Der Schutz der Einzelangaben ist also bei korrekter Anwendung des Verfahrens sichergestellt.
3. Akzeptanz der Ergebnisse	Bei der Ausgestaltung der Übergangsmatrix wird berücksichtigt, dass die Veränderungen (unter Berücksichtigung von Punkt 2) klein genug sind, um bei einem Großteil der Nutzenden auf Akzeptanz zu stoßen.	Je nach Größe der betrachteten (Sub-)Population und der Detailtiefe der betrachteten Merkmale werden die Sperrungen von einigen Nutzenden als zu restriktiv betrachtet.
4. Ergebnisse sind konsistent	Logisch identische Tabellenfelder werden immer mit dem gleichen Wert überlagert.	Sperrmuster werden auch tabellenübergreifend umgesetzt, sodass Veränderungen bei korrekter Umsetzung des Zellsperrverfahrens konsistent sind. Bei häufig genutzten Statistiken und vielen Ergebnistabellen können übersehene Zusammenhänge zwischen Tabellenfeldern allerdings zu Fehlern führen.
5. Ergebnisse sind additiv	Durch die unabhängige Überlagerung der einzelnen Fallzahlen sind geheim gehaltene Tabellen nicht additiv.[1]	Additivität bleibt in Tabellen bestehen, sofern keine Sperrungen erfolgen.

Ziel voll erreicht Ziel teilweise erreicht Ziel nicht erreicht

1 Die Additivität könnte durch einen Algorithmus wiederhergestellt werden, der die überlagerten Werte so verändert, dass sich die Tabelleninnenfelder wieder zu den Randfeldern addieren. Dieses Vorgehen hätte aber zwei entscheidende Nachteile: Zum einen erfordert die Herstellung der Additivität zusätzliche Rechenleistung und verlängert die für die Tabellenerstellung benötigte Laufzeit, zum anderen wären die Ergebnisse danach nicht mehr konsistent, was die Qualität der Daten entscheidend beeinträchtigen würde.

3

Verhältniszahlen

Verhältniszahlen sind mathematische Beziehungen zwischen statistischen Größen, die einen sinnvollen Zusammenhang darstellen. Generell können drei Kategorien von Verhältniszahlen unterschieden werden (Bourier, 2011):

1. Gliederungszahlen: Diese setzen eine Teilgröße in Beziehung zur Gesamtgröße und werden oft als Anteilswerte bezeichnet, wie bei Geschlechterquoten.
2. Beziehungszahlen: Hier werden unterschiedliche Größen miteinander in Beziehung gesetzt, die in einem sachlichen Zusammenhang stehen, wie das Bruttoinlandsprodukt je Einwohner/-in.
3. Messzahlen: Diese setzen inhaltlich ähnliche Größen in Beziehung, jedoch zu unterschiedlichen Zeitpunkten, Zeiträumen oder Regionen, wie die Veränderung des Bruttoinlandsprodukts im Vergleich zum Vorjahr.

In der wissenschaftlichen Forschung ist die Berechnung von Verhältniszahlen eine gängige Methode. Dabei ist es wichtig, die geltenden Geheimhaltungsvorschriften und -regelungen zu beachten, insbesondere auch im Rahmen von Forschungsprojekten in den Forschungsdatenzentren. Eine grundlegende Anforderung besteht vor diesem Hintergrund darin, die Geheimhaltung gemäß den fachspezifischen Konzepten sicherzustellen, um ein konsistentes Vorgehen über verschiedene Analysen und Veröffentlichungen hinweg zu gewährleisten. Dies kann auch spezielle statistische Regelungen für den Umgang mit Verhältniszahlen erforderlich machen, die von den Forschungsdatenzentren wie von ihren Nutzenden zu berücksichtigen sind.

Für Nutzende der Forschungsdatenzentren ist es entscheidend, die Methodik der Geheimhaltung und deren Auswirkungen zu verstehen, um die Qualität der berechneten Verhältniszahlen selbst einschätzen zu können. Der folgende Abschnitt erläutert die Auswirkungen der Cell-Key-Methode auf Verhältniszahlen sowie auf deren Aussagekraft.

3.1 Auswirkungen auf Verhältniszahlen

Basis für die Geheimhaltung von Verhältniszahlen mit der Cell-Key-Methode ist die Veränderung der Zähler und Nenner entsprechend ihrer Cell Keys. Dadurch kann es vor allem bei kleinen Fallzahlen passieren, dass sich das errechnete Verhältnis deutlich vom Wert der nicht überlagerten Verhältniszahl unterscheidet. Diese Ungenauigkeiten verringern sich mit größeren Fallzahlen. Aus diesem Grund wird von einer Berechnung von Verhältniszahlen abgeraten, wenn diese auf nur wenigen Fällen beruhen.

3.2 Umgang mit Verhältniszahlen

Häufig werden Verhältniszahlen aus den überlagerten Fallzahlen von Zähler und Nenner berechnet, was hier als „A-posteriori-Verhältniszahl“ bezeichnet wird.[3] Daher kann es bei datenverändernden Geheimhaltungsverfahren wie der Cell-Key-Methode zu unerwünschten Effekten kommen:

- Ungenauigkeit: Bei der Anwendung der Cell-Key-Methode kann die Veränderungsrichtung der Fallzahlen zufällig stark gegenläufig sein, das heißt Zähler und Nenner werden um einen hohen positiven beziehungsweise negativen Wert (oder umgekehrt) verändert. Das kann insbesondere bei kleinen Fallzahlen zu erheblichen Abweichungen zwischen der ursprünglichen und der A-posteriori-Verhältniszahl führen. Dieser Effekt verringert sich jedoch mit steigenden Fallzahlen.
- Unplausibilität: Ein weiterer unerwünschter Effekt kann auftreten, wenn zum Beispiel der ursprünglich kleinere Zähler nach der Veränderung größer ist als der veränderte Nenner. In diesem Fall würden sich A-posteriori-Anteilswerte über 100 % ergeben.
- Veränderung der Aussage: Besonders bei der Analyse von Zeitreihen kann die Anwendung der Cell-Key-Methode zu Trendverzerrungen oder sogar zu einer Trendumkehr führen (siehe Kapitel 4).

Während die Überlagerung von Fallzahlen eine feste Varianz aufweist, stellen Enderle und andere (2018) fest, dass dies nicht auf Verhältniswerte zutrifft, die auf Basis zweier überlagerter Werte berechnet wurden. Für das Ausmaß der Abweichung des tatsächlichen Verhältniswerts $R := X/Y$ zum überlagerten Verhältniswert $\hat{R} := \hat{X}/\hat{Y}$ spielen die Originalfallzahlen, aus denen das Verhältnis berechnet wird, eine große Rolle: je größer die Fallzahlen X und Y, desto geringer die Abweichung des Verhältniswerts. Zur Veranschaulichung der Problematik kleiner Fallzahlen nennen Enderle und andere (2018) folgendes Beispiel: Die relativ kleinen Originalfallzahlen $x = 4$ und $y = 4$ werden zueinander ins Verhältnis gesetzt. Für den tatsächlichen Verhältniswert gilt damit: $R = 4/4$, also 100 %. Bei einer Maximalabweichung von $D = 3$ könnten die überlagerten Fallzahlen die Werte $\hat{x} = x + D = 7$ und $\hat{y} = y - D = 1$ annehmen. Der überlagerte Verhältniswert wäre dann $R = 7/1$, also 700 %.

3.3 Beurteilung der Qualität eines Anteilswertes

Die Qualität eines Anteilswertes muss von Wissenschaftlerinnen und Wissenschaftlern bewertet werden können. Dafür sind Informationen über die Verteilung der zufälligen Abweichungen von der Originalgröße erforderlich, wobei jedoch die Details dieser Abweichungen nicht offengelegt werden dürfen. Um die Qualität von Anteilswerten zu beurteilen kann die relative Standardabweichung als ein Maß für die Streuung um die Originalfallzahl verwendet werden.[4]

Die relative Standardabweichung für Anteilswerte oder für alle Verhältniszahlen, die als Quotienten aus veränderten Fallzahlen im Zähler und Nenner gebildet werden, wird wie folgt berechnet:

Angenommen, q_{xy} repräsentiert die Wahrscheinlichkeit, dass einem veränderten Anteilswert $\tilde{v} := \widetilde{\left(\frac{x}{y}\right)}$ der Originalzähler $x \in \mathcal{D}_x$ und der Originalnenner $y \in \mathcal{D}_y$ zugrunde liegen, und $d(v)_{xy}$ repräsentiert die Abweichungen zwischen dem veränderten Anteilswert $\tilde{v}$ und den möglichen originalen Anteilswerten v, dann kann die (absolute) Standardabweichung für den veränderten Anteilswert $\tilde{v}$ wie folgt approximiert werden:

3 Weiterführende Vorschläge für Techniken zur Anwendung stochastischer Überlagerung bei Verhältniszahlen, die als Quotient aus zwei Wertsummen gebildet werden, finden sich in Gießing (2013).

4 Anstelle der absoluten Standardabweichung wird die relative Standardabweichung als Gütemaß herangezogen, da diese die zufällige Streuung um die Originalfallzahl ins Verhältnis zur Größe des Originalwertes setzt.

$$\sigma(\tilde{v}) = \sqrt{\sum_{\substack{x\in\mathcal{D}_x \\ y\in\mathcal{D}_y}} q_{xy}\cdot d(v)_{xy}^2}$$

Der entsprechende Relative Root Mean Square Error (RRMSE) oder die approximierte relative Standardabweichung (auch Variationskoeffizient genannt) für den veränderten Anteilswert $\tilde{v}$ ergibt sich dann gemäß

$$k(\tilde{v}) = \frac{\sigma(\tilde{v})}{\tilde{v}}.$$

Für Mittelwerte oder Verhältniswerte, bei denen die Wertsumme im Zähler und/oder Nenner (gegebenenfalls verändert) in die Berechnung einfließt, erfordert das beschriebene Verfahren einige Anpassungen. Eine alternative Vorgehensweise besteht darin, die relative Standardabweichung von v mithilfe der Übergangswahrscheinlichkeiten zu bestimmen.

Für die Hochschulstatistik wurden die Auswirkungen unterschiedlicher Parametrisierungen auf die Qualität von Verhältniszahlen untersucht. Das Ergebnis zeigt, dass die Wahl von Bleibewahrscheinlichkeiten und maximaler Abweichung von der Originalfallzahl nicht die entscheidenden Faktoren für die Qualität von Verhältniszahlen darstellen. Vielmehr hängt die Qualität vor allem von der Größe der zugrunde liegenden Fallzahlen ab, die in die Berechnung einfließen. Demnach beeinträchtigen insbesondere kleine Fallzahlen die Qualität und Aussagekraft der Verhältniszahlen. Erst ab einer bestimmten Größe der Basiszahlen im Zähler und Nenner kann eine ausreichende statistische Aussagekraft gewährleistet werden (Enderle/Vollmar, 2019).

3.4 Empfehlungen bei der Berechnung von Verhältniszahlen

Ohne Kenntnis der unveränderten Verhältniszahlen ist es für Nutzende nicht möglich, die Qualität einer Verhältniszahl zu beurteilen. Die Höhe der relativen Standardabweichung oder des RRMSE hängt in erster Linie von der Größe der einbezogenen Fallzahlen ab. Daher wird allgemein empfohlen, Verhältniszahlen nur auf Basis ausreichend hoher Fallzahlen im Zähler und Nenner zu berechnen, denn die Varianz geht in diesen Fällen ohnehin gegen Null.

Sollte die Berechnung einzelner Verhältniszahlen auf Basis geringer Fallzahlen für eine Forschungsfrage unerlässlich sein, können Nutzende der Forschungsdatenzentren ihren betreuenden FDZ-Standort um Unterstützung bitten.

4

Zeitreihen

Zeitreihen sind wertvolle Instrumente, um zeitliche Verläufe und Entwicklungen darzustellen. Hierbei werden Kennzahlen auf der Basis von Zeitpunkten oder Zeiträumen berechnet. Dies ermöglicht die Analyse von absoluten oder relativen Veränderungen über die Zeit hinweg.

Wie bei der Berechnung von Verhältniszahlen sind auch bei der Betrachtung von Zeitreihen die geltenden Geheimhaltungsvorschriften und -regelungen der Fachstatistik zu beachten.

4.1 Auswirkungen auf Zeitreihen

Die Anwendung der Cell-Key-Methode kann negative Auswirkungen auf die Analyse von Zeitreihen haben. Das gilt sowohl beim Vergleich von zwei mittels Cell-Key-Methode geheim gehaltenen Erhebungswellen als auch bei der Betrachtung von zwei Erhebungswellen mit unterschiedlicher Geheimhaltung, also beim Vergleich der Erhebungsjahre vor und nach Einführung der Cell-Key-Methode. Besonders bei sehr geringen Unterschieden in den unveränderten Fallzahlen der beiden verglichenen Beobachtungszeiträume kann es bei gegenläufiger Veränderung der betrachteten Werte vorkommen, dass Unterschiede zwischen den Erhebungswellen vergrößert oder verkleinert werden. Im Extremfall ist sogar eine Trendumkehr möglich. Sehr schwache Veränderungen im Zeitverlauf sind bei Anwendung der Cell-Key-Methode daher mit Vorsicht zu interpretieren.

4.2 Umgang mit Zeitreihen

Zeitreihen basieren auf Daten aus verschiedenen Zeiträumen und werden verwendet, um die zeitliche Entwicklung von Sachverhalten zu analysieren. Die einzelnen Datenpunkte zu den verschiedenen Berichtszeiträumen bilden die Basis für entsprechende Betrachtungen.

Die Bildung von Differenzen aus veränderten Datenpunkten kann zu weniger sicheren Ergebnissen führen. Das gilt insbesondere dann, wenn eine Differenz anhand kleiner Fallzahlen berechnet wird.

Bei der Berechnung von Differenzen sind bei der Geheimhaltung mit der Cell-Key-Methode spezifische Herausforderungen zu berücksichtigen:

I. Umgang mit Differenzen, wenn beide Datenpunkte aus CKM-Datenbeständen stammen

Bei der Differenz zwischen zwei zufällig veränderten Datenpunkten können ähnliche Herausforderungen auftreten wie bei Saldierungen (siehe Abschnitt 4.3). Dies trifft insbesondere bei kleinen Fallzahlen zu und wenn die einzelnen Punkte starke, gegenläufige Veränderungen aufweisen. In der Konsequenz kann die Anwendung der Cell-Key-Methode theoretisch Veränderungen in der Trendstärke und sogar eine Änderung des Trendvorzeichens verursachen.

Für diese Herausforderungen gibt es (statistikspezifische) Lösungsansätze. Im Zensus 2022 wird eine Methode gewählt, die einen Vorzeichenwechsel vermeidet. In der Bevölkerungsstatistik und der Hochschulstatistik hingegen wird die Differenz aus den beiden überlagerten Datenpunkten gebildet. Aus methodischer Sicht führt dies zu einer Verdopplung der in den CKM-Parametern vorgegebenen Varianz. Wenn X_1 und X_2 zwei veränderte Datenpunkte mit identischer, als CKM-Parameter festgelegter Varianz V sind, gilt für ihre Differenz:

$$Var\,(X_1 - X_2) = Var\,(X_1) + Var\,(X_2) = 2V$$

Dies erhöht im Vergleich zu den einzelnen absoluten Fallzahlen das Risiko einer größeren Abweichung von der Originaldifferenz. Basieren die Zeitreihen (oder die einzelnen Datenpunkte) jedoch auf ausreichend großen Fallzahlen, wird der Effekt einer Verzerrung der Trendstärke vernachlässigbar.

Um das Risiko einer wesentlichen Veränderung der Trendstärke zu minimieren, empfehlen die Forschungsdatenzentren, in wissenschaftlichen Veröffentlichungen Analysen auf Basis von Zeitreihen nur auf Basis einer ausreichend hohen Fallzahl durchzuführen. Wie hoch diese Fallzahlgrenze ist, hängt von der genutzten Statistik ab sowie von den betrachteten Merkmalen. Der für die jeweilige Statistik fachlich zuständige FDZ-Standort kann hierzu beratend unterstützen.

II. Umgang mit Differenzen, wenn die Datenpunkte aus unterschiedlich geheim gehaltenen Datenbeständen stammen

Ändert sich das Geheimhaltungsverfahren zwischen zwei Berichtszeitpunkten, beispielsweise durch den Wechsel von der Zellsperrung zur Cell-Key-Methode, führt dies zu einem methodischen Bruch in der Zeitreihe. Die Forschungsdatenzentren haben für diesen Fall geregelt, dass an dem Punkt des Bruchs, also dort, wo der Wechsel des Geheimhaltungsverfahrens erfolgt, keine Differenzen berechnet werden dürfen. Dies hat zur Konsequenz, dass die Zeitreihe am Bruchpunkt unterbrochen ist. Vergleiche zwischen den Berichtszeitpunkten der Umstellung sind daher nicht möglich. Einige Fachstatistiken, wie der Zensus, haben jedoch entschieden, frühere Wellen ihrer Statistik nachträglich ebenfalls mit der Cell-Key-Methode geheim zu halten. Entsprechende Vergleiche anhand der beiden mit der Cell-Key-Methode geheim gehaltenen Datenbestände sind dann wieder möglich. Die Information, ob das für eine bestimmte Statistik durchgeführt wurde, findet sich in den entsprechenden von den Forschungsdatenzentren bereitgestellten Metadatenreports.

4.3 Umgang mit Saldierungen

Der Umgang mit Saldierungen ist in Bezug auf die grundlegende Problematik sehr ähnlich dem Umgang mit Zeitreihen. Wird ein Saldo, also die Differenz zweier veränderter Fallzahlen, berechnet (beispielsweise der Wanderungssaldo in der Bevölkerungsstatistik), so führt die Verknüpfung zweier stochastischer Größen zu einer höheren Unsicherheit des Ergebnisses. Dies geschieht, weil die Varianz der Differenz im Vergleich zur vorhandenen Varianz bei der Veränderung einzelner Fallzahlen verdoppelt wird (siehe Abschnitt 4.2). Dadurch kann die Aussagekraft des Saldos bei kleinen Fallzahlen verzerrt werden. Zudem besteht die theoretische Möglichkeit, dass bei Anwendung der Cell-Key-Methode das Vorzeichen des Saldos wechselt, wenn beide Originalfallzahlen gegenläufig verändert werden und die Fallzahlen sowohl klein sind als auch nahe beieinanderliegen. Der relative Effekt solcher Verzerrungen aufgrund möglicher großer Abweichungen von der ursprünglichen Differenz nimmt jedoch ab, je größer die zugrunde liegenden Fallzahlen sind – ähnlich wie bei Verhältniszahlen und Zeitreihendifferenzen.

Die Forschungsdatenzentren empfehlen daher auch bei Saldierungen, entsprechende Auswertungen nur auf Basis ausreichend hoher Fallzahlen vorzunehmen. Bei vielen kleinen Fallzahlen sollte die Auswertung nach Möglichkeit auf einer höheren regionalen oder fachlichen Aggregationsebene (zum Beispiel Landkreise statt Gemeindeebene oder Wirtschaftszweig-4-Steller statt -5-Steller) erfolgen, ebenso sollten schwach besetzte Kategorien gruppiert werden.

5

Fazit

Im Vergleich mit den traditionellen Geheimhaltungsverfahren, allen voran der weit verbreiteten Zellsperrung, weist die Cell-Key-Methode mit Blick auf die Abwägung zwischen Informationsverlust und Aufdeckungsrisiko einige Vorteile auf. Allerdings hat dieser Beitrag ausführlich dargestellt, dass die Datennutzenden sowie die Forschungsdatenzentren auch neue Besonderheiten bei der Auswertung und Interpretation der mit der Cell-Key-Methode geheim gehaltenen Ergebnisse berücksichtigen müssen.

Da die Cell-Key-Methode für die Geheimhaltung von Fallzahltabellen ausgelegt ist, kann es bei darüber hinausgehenden Analysen auf Basis von Verhältniszahlen, Zeitreihen und Saldierungen zu Problemen kommen. Diese sind vermeidbar, sofern Auswertungen immer auf Basis ausreichend großer Fallzahlen vorgenommen werden. Sollten bei der Nutzung in den Forschungsdatenzentren Probleme oder Unsicherheiten entstehen, können Nutzende sich jederzeit an ihren betreuenden FDZ-Standort wenden.

LITERATURVERZEICHNIS

Bourier, Günther. *Beschreibende Statistik. Praxisorientierte Einführung – Mit Aufgaben und Lösungen*. Wiesbaden 2011, Seite 19 ff. DOI: 10.1007/978-3-8349-6556-1

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. DOI: 10.1007/978-3-319-99771-1_3

Enderle, Tobias/Vollmar, Meike. *Geheimhaltung in der Hochschulstatistik.* In: WISTA Wirtschaft und Statistik. Ausgabe 6/2019, Seite 87 ff.

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: unece.org

Giessing, Sarah. *What shall we do with the ratios?* Work session on statistical data confidentiality. Supporting paper. Ottawa 2013. [Zugriff am 7. Mai 2024]. Verfügbar unter: unece.org

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

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: 2011.isiproceedings.org

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).

Setzer, Stefanie/Rohde, Johannes/Güttgemanns, Volker/Rothe, Patrick. *Die Cell-Key-Methode in den Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder - Teil 1: Vorstellung des neuen Geheimhaltungsverfahrens*. In: WISTA Wirtschaft und Statistik. Ausgabe 3/2024, Seite 31 ff.

Herausgeber
Statistisches Bundesamt (Destatis), Wiesbaden

Schriftleitung
Dr. Daniel Vorgrimler
Redaktion: Ellen Römer

Ihr Kontakt zu uns
www.destatis.de/kontakt

Erscheinungsfolge
zweimonatlich, erschienen im Juni 2024
Ältere Ausgaben finden Sie unter www.destatis.de sowie in der Statistischen Bibliothek.

Artikelnummer: 1010200-24003-4, ISSN 1619-2907