Otto-Friedrich-Universität Bamberg
Lehrstuhl für Statistik und Ökonometrie
Standort: Bamberg
Sommersemester 2022
Vorlesung: Blockseminar Survey-Methodik
Prüfer: Dr. Sara Bleniger

Statistische Geheimhaltung: Cell Key Methode

Joshua Simon
joshua-guenter.simon@stud.uni-bamberg.de
Master Survey Statistik, 4. Fachsemester
Matrikelnummer: 2032411
25. Juli 2022

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Listings

1. Einführung

Die amtliche Statistik sorgt mit einer Vielzahl an Veröffentlichungen für die Bereitstellung von aufbereiteten statistischen Informationen. Damit geht sie dem Ziel nach, Bürgern, Institutionen und anderen gesellschaftlichen Einrichtungen eine Datengrundlage für die Entscheidungsfindung zu bieten. Weiter dient die amtliche Statistik auch der Politik und der Wissenschaft als Datenquelle. Das Sammeln und Erheben dieser Daten stellt in vielen Fällen einen Eingriff auf das Recht der informationellen Selbstbestimmung für Personen und Entitäten dar. Dieses Recht ist das Fundament des mordernen Datenschutzes und wird über Artikel 2 des Grundgesetztes abgedeckt. Es steht außer Frage, dass dieses Recht besonders schützenswert ist. Demnach steht auch die amtliche Statistik in der Pflicht dieser Verantwortung nachzukommen. Konkret manifestiert sich das Einhalten dieser Plficht in dem sog. Statistikgeheimnis. Aus dem Bundestatistikgesetzt lässt sich hierzu der folgende Absatz aufgreifen (§16 Abs. 1 Satz 1 BStatG):
„Einzelangaben über persönliche und sachliche Verhältnisse, die für eine Bundesstatistik gemacht werden, sind von den Amtsträgern und für den öffentlichen Dienst besonders Verpflichteten, die mit der Durchführung von Bundesstatistiken betraut sind, geheim zu halten, soweit durch besondere Rechtsvorschrift nichts anderes bestimmt ist.“
Konkret möchte man mit der statistischen Geheimhaltung einen Schutz für einzelne Person und Entitäten vor der Offenlegung ihrer sensitven Daten bieten. Dies dient im Weiteren auch der Aufrechterhaltung des Vertrauensverhältnisses zwischen den Befragten und den statistischen Ämtern und erhebenden Einrichtungen. Dies gewährleistet abermals die Zuverlässigkeit der Angaben und der Berichtswilligkeit der Befragten. Die vorausgegangenen Punkte werden in einer Begrüdung zum BStatG erwähnt [Nickl, 2019]. Ausnahmen von einer Geheimhaltung bestehen nur in Ausnahmefällen, z.B. wenn eine explizite Einwillung zur Veröffentlichung durch den Befragten vorliegt oder wenn sich die Informationen aus allgemein zugänglichen Quellen von öffentlichen Stellen beziehen. Auch die inner-beördliche Übermittlung, Methodenentwicklung, Planungs- und Forschungszwecke werden über das BStatG geregelt.
Im weiteren Verlauf dieser Arbeit sollen Geheimhaltungsverfahren und Geheimhaltungsregeln präsentiert werden, die im Einzelnen die statistische Geheimhaltung gewährleisten und damit dem Statistikgeheimnis der amtlichen Statistik nachkommen. Besondere Beachtung wird dabei der Cell Key Methode (CKM) geschenkt. Dieses Verfahren bietet ein Ansatz, welcher gegenüber anderen Verfahren gut zu Implementieren und zu Automatisieren ist. Gerade dieser Punkt ist in einem immer

weiter werdenden technolgischen Umfeld nicht außer Acht zu lassen.

2. Etablierte Geheimhaltungsverfahren

In diesem Abschnitt sollen zunächst die grundlegenden Begrifflichkeiten für Geheimhaltungsverfahren und Geheimhaltungsregeln beschrieben werden.

2.1. Methodische Grundlagen

Grundsätzlich gibt es zwei sich unterscheidenede methodische Ansätze, die bei der Geheimhaltung zum Tragen kommen können [Nickl, 2019]. Zum einen existieren *informationsreduzierende Methoden*. In der Gattung dieser Verfahren werden durch Aggregation oder Sperrrung kritische Kategorien oder Werte die Aufdeckungsrisiken verhindert. Eine Aggregation meint in diesem Fall das Zusammenfassen zu übergeordneteten Positionen, z.B. durch Summation kleinerer Positionen. Bei einer Sperrung ist auch oft von einer Löschung die Rede. Hier werden gezielt einzelne Werte identifiziert und aus der Tabelle entfernt. Als Kontrast stehen *datenverändernde Methoden* gegenüber. Hier werden durch gezielte Veränderungen der Daten - beispielsweise durch Runden oder Zufallsüberlagerungen - kritische Werte verfälscht, was auch für eine erfolgreiche statistische Geheimhaltung sorgen kann.
Weiter differenziert man Geheimhaltungsverfahren auch nach dem Zeitpunkt ihrer Durchführung. Die Daten können bereits vor der Tabellierung mit einer Geheimhaltung versehen werden. Man spricht hier von *pre-tabulare Verfahren* [Rothe, 2015-5]. Diese Verfahren werden als Anonymisierung bezeichnet, da die Daten im Vorfeld so verändert werden, dass keine kritischen Ergebnisse resultieren. Oftmals ist diese Art von Geheimhaltung aber nicht ausreichend, weshalb weitere Verfahren im Anschluss angewandt werden müssen. Man spricht nun von *post-tabulare Verfahren* [Rothe, 2015-5]. Ihre Mechanismen werden auf die fertig tabellierten Daten angewandt.

2.2. Statistische Tabellen

Gegenstand der Geheimhaltung stellen in dieser Arbeit statistische Tabellen dar. Ein Großteil der Veröffentlichungen der amtlichen Statistik sind selbst - oder beinhalten - statistische Tabellen, welche aus den amtlichen Daten abgeleitet werden. Maßgebend für die Anwendung eines Geheimhaltungsverfahrens ist die Art der zu veröffentlichenden Tabelle, die vorliegt. Man unterscheidet im allgemeinen zwischen *Häufigkeitstabllen* und *Wertetabellen* [Nickl, 2019]. Erstere stellen Häufigkeiten oder Fallzahlen dar, z.B. die Anzahl von Frauen und Männern innerhalb einer Universitätät. Wertetabellen hingegen stellen Wertesummen wie Umsätze dar. Demnach sind sie häufig in Wirtschaftsstatistiken anzutreffen [Rothe, 2015-8]. Diese unterschiedlichen Kontexte, in denen die Zahlen dieser Tabellen interpretiert werden können, fordern eine natürliche Unterscheidung innerhalb der Geheimhaltung. Es folgen zwei einfache Beispiel für diese Tabellentypen mit rein fiktiven Ausprägungen.

Studienfach	**männlich**	**weiblich**	**insgesamt**
Bauingenieurwesen	4	3	7
Informatik	9	12	21
Medizin	4	1	5
Survey Statistik	10	10	20
Gesamt	27	26	53

Tabelle 1: Beispiel für eine Häufigkeitstabelle

Brauerei	**Mährs Bräu**	**Schinkerla**	**Käsmann**	**Gesamt**
Umsatz	600 000	50 000	250 000	900 000

Tabelle 2: Beispiel für eine Wertetabelle

2.3. Bewährte Ansätze

Nachdem nun die grundlegenden Begriffe im Zusammenhang mit der Geheimhaltung von statistischen Tabellen erläutert wurden, beschäftigt sich dieser Abschnitt mit bewährten Ansätzen um diese Geheimhaltung durchzuführen. Zunächst werden die Mechanismen für eine pre-tabulare Geheimhaltung - also eine Anonymisierung - beschrieben. Gegenstand der Beachtung sind hier immer die Einzeldatensätze. Es existieren verschiedene Stufen der Anonymität. Diese werden in [Rothe, 2015-5] wie folgt dargestellt:

- **formale Anonymisierung:** Diese Art der Anonymisierung wird intern in den statistischen Landesämtern verwendet. Durch Entfernung aller direkten Identifikatoren [Nickl, 2019], wie z.B. Name, Adresse oder Matrikelnummer, bleibt der eigentliche Informationsgehalt für die statistischten Auswertungen hoch.

- **faktische Anonymisierung:** Für die unabhängige wissschaftliche Forschung werden die Daten so bearbeitet, dass unter realistischen Rahmenbedingungen keine erfolgreiche Identifikation möglich ist [Nickl, 2019].

- **absolute Anonymisierung:** Diese Variante bietet den höchsten Informationsverlust, stellt im Gegenzug aber sicher, dass unter keinen Umständen eine

Reidentifikation vorgenommen werden kann [Nickl, 2019]. Damit ist dieser Ansatz zu wählen, wenn die Tabellen an die breite Öffentlichkeit weitergeben werden sollen.

Bei allen drei Optionen ist das darunterliegende Vorgehen identisch. Im ersten Schritt wird immer die jeweilige Anonymisierung auf die Einzeldatensätze angewandt. Anschließend werden die statistischen Tabellen generiert. Für weitere Einzelheiten sei an dieser Stelle auf [Rothe, 2015-5] verwiesen.
Im Weiteren wird speziell auf die post-tabularen Verfahren für Häufigkeits- und Wertetabelle eingangen. Am Anfangs eines jeden Geheimhaltungsverfahrens stehen die sog. *Geheimhaltungsregel*, die eine Identifizierung kritischer Fälle erlauben. Eine Standardansatz in der amtlichen Statistik stellt hier die *Mindestfallzahlregel* [Rothe, 2015-5] dar. Diese Regel lässt sich auf Häufigkeitstabelle anwenden, indem die kritischen Fälle mit einem zuvor festgelegten Wert n verglichen werden. Meist wird hier $n = 3$ gewählt [Rothe, 2015-5]. Ist der betrachtete Tabellenwert kleiner als der Wert n, so ist dieser Tabellenwert geheimzuhalten. Für Wertetabellen hingegen lässt sich beispielsweise die *p-% Regel* anwenden [Rothe, 2015-8]. Dieser Ansatz besagt, dass ein Tabellenwert x geheimzuhalten ist, wenn

$$x - x_2 - x_1 < \frac{p}{100} \cdot x_1 \tag{1}$$

gilt [Rothe, 2015-8]. Wobei in (1) x_1 der größte und x_2 der zweitgrößte Beitrag ist. In anderen Worten lässt sich sagen, dass der Tabellenwert x genau dann geheimzuhalten ist, wenn die Differenz zwischen x und x_1, x_2 nicht mindestens p% vom größten Beitrag x_1 beträgt.
Um letztlich auch die Geheimhaltung durchzuführen wird unter diesen Regeln oft auf das Verfahren der *Zellsperrung* zurückgegriffen [Rothe, 2015-5]. Hierbei werden die in einem ersten Schritt mit der Geheimhaltungsregel identifizierten Fälle durch einen Punkt $\cdot$ ersetzt. Bei dieser Sperrung der kritischen Werte spricht man von der *Primärsperrung*. Um aber auch Rückrechnungen unter Zuhilfenahme von Rand- oder Zwischensummen zu verhindern, wird eine *Sekundärsperrung* angewandt. Diese sperrt neben dem bereits durch die Primärsperrung unkenntlichen gemachten Zellwerts auch jeweils eine Zellwert in derselben Zeile, in derselben Spalte sowie dasjenige Tabellenfeld, in dem die Zeile und die Spalte der beiden zuvor genannten Felder aufeinander treffen [Rothe, 2015-5]. Es wird schnell klar, das unter Anwendung dieses Verfahres ein hoher Informationsverlust im Gunste der Geheimhaltung entstehen kann. Um dieses Verlust möglichst gering zu halten, sollten eher Tabelleninnenfelder und keine Summenfelder gesperrt werden [Rothe, 2015-5]. Die Zellsperrung stellt

nach Anwendung ein sicheres Verfahren dar, welches allerdings durch die oftmals mit hohem manuellen Aufwand verundenen Sekundärsperrung eher unfelxibel ist. Für ein anschauliches Beispiel sei an dieser Stelle auf [Rothe, 2015-5] verwiesen.

3. Cell Key Methode

Die im vorherigen Kapitel beschrieben Geheimhaltungsverfahren müssen in der Regel - zumindest bis zu einem gewissen Grad - manuell durchgeführt werden und eine Automatisierung ist eher unfelxibel. Mit der *Cell Key Methode (CKM)* wird ein Geheimhaltungsverfahren präsentiert, welches gut zu automatisieren und vergleichsweise einfach zu implementieren ist. Die Cell Key Methode ist auch als *ABS-Verfahren* bekannt. Der Name stammt von der schöpfenden Institutuion des Verfahrens, dem Australian Bureau of Statistics, ab. Durch die Verwendung von zufallsbasierten Additionen, den sog. Überlagerungen, werden Datenwerte verschleiert. Der Ermittlung einer solchen zufallsbasierten Addition liegt eine einmalig festzulegende Wahrscheinlichkeitsverteilung mit den möglichen Überlagerungswerten zugrunde [Enderle, 2019]. Für die Bestimmung dieser Überlagerungen wird ein deterministischer Mechaniusmus eingesetzt. Dieser nutzt den original Zellwert und den sog. Cell-Key um aus der Verteilung der Überlagerungswerte eine eindeutige Überlagerung zu ziehen [Enderle, 2019]. Eine mit dem CKM Verfahren geheimgehaltene Tabelle veröffentlicht also die Summe aus Originalwerten und Überlagerungen. Die CKM zählt damit zu den datenveränderndeden Verfahren.

3.1. Verfahrensparameter und Überlagerungsmatrix

An statistische Geheimhaltungsverfahren, insbesondere den datenverändernden Verfahren, werde bestimmte Anforderungen gestellt. Für die Cell Key Methode werden in der amtlichen Statistik gewisse stochastische Eigenschfaten gefordert, um die Qualität und Nachvollziehbarkeit der Ergebnisse zu sichern. Zu diesen Eigenschaften zählt einerseits die Unverzerrtheit der Überlagerungen [Enderle, 2019]. Damit meint man, dass der Überlagerungswert, welcher zu den Originalwerten addiert wird, im Mittel gleich Null ist. Es soll also die Erwartungstreue $E(z) = 0$ gelten. Darüber hinaus fodert man ebenfalls eine konstante Streuung der Verteilung der Überlagerungen [Enderle, 2019]. Es soll die Varianz erhalten bleiben, also $Var(z) = s^2$ gelten. Um diese beiden Eigenschaften zu kokretisieren, werden Verfahrensparameter eingeführt. Diese dienen weiter auch dazu das Geheimhaltungsverfahren - und damit die Überlagerungen - an verschiedene Kontexte anzupassen. Zu diesen Verfahrensparameter zählen nach [Höhne, 2019]:

- Eine boolsche Variable, die angibt, ob Originalwerte 1 und 2 geheimgehalten werden sollen.
- Der Anteil p_0 der nicht zu überlagernden Originalwerte.
- Die Maximalüberlagerung d.
- Die Standardabweichung der Überlagerungsbeiträge s.

Gestand des Interesses sind demnach Zufallsfunktionen z, die eine bedingte Wahrscheinlichkeitsverteilung auf die Originalwerte i mit Zielhäufigkeit j darstellen. Man sucht also für jeden Originalwert $i = \{0, 1, 2, \ldots, n\}$ eine Wahrscheinlichkeitsverteilung der Form $z = p_i$ mit den Wahrscheinlichkeiten für die Übergänge v_i hin zu den Zielhäufigkeiten j [Enderle, 2019]. Diese Wahrscheinlichkeiten bilden ein nichtlineares Gleichungssystem, in welchem die zuvor genannten stochastischen Eigenschaften als Nebenbedinungen eingehen. Diese lassen sich nun als

$$E(z) = \sum_{i=-d}^{d} p_i v_i = 0 \tag{2}$$

$$Var(z) = \sum_{i=-d}^{d} p_i v_i^2 = s^2 \tag{3}$$

$$\sum_{i=-d}^{d} p_i = 1 \tag{4}$$

schreiben [Höhne, 2019]. Dabei wird in der letzten Zeile (4) noch gefordert, dass die Summe der Übergangswahrscheinlichkeiten für einen Originalwert gleich 1 ist. Die Lösung dieses Problems lässt sich in Matrixform notieren. Man spricht hier von der sog. Überlagerungsmatrix mit Zeilen i und Spalten j, welche zur Bestimmung der Überlagerungsbeiträge j zu den Originalwerten i verwendet werden kann. Für weitere mathematische Details und Lösungsansätze sei an dieser Stelle auf [Giessing, 2016] verwiesen. In [Höhne, 2019] wird die Lösung eines solchen Gleichungssystems für die Verfahrensparameter $p_0 = 0{,}5 = 50\%$, $d = 4$ und $s = 2{,}25$ gezeigt. Diese Überlagerungsmatrix ist in Abbildung 1 in Form einer Heatmap zu sehen. Der *Python* Code zum Erstellen dieser Grafik ist in Anhang A zu finden.

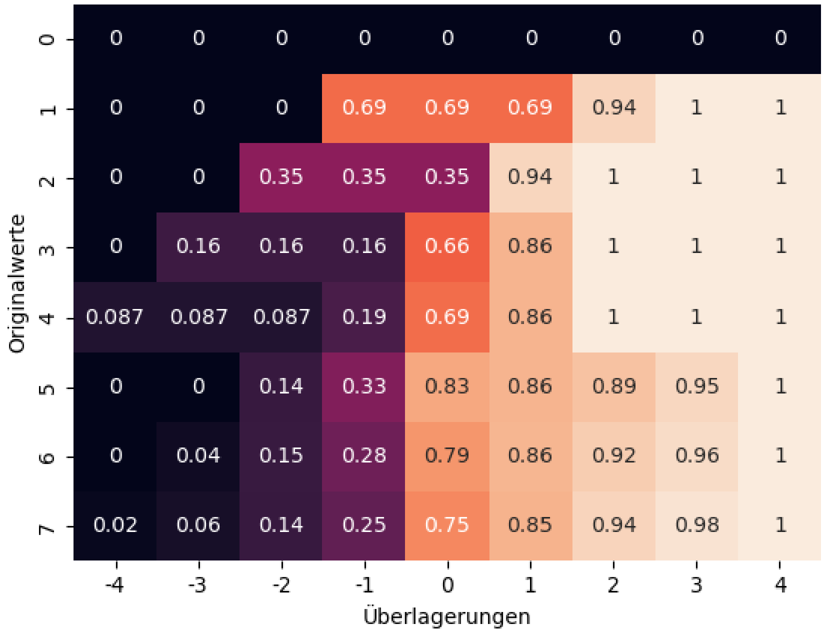

Abbildung 1: Beispiel für eine Überlagerungsmatrix mit Übergangswahrscheinlichkeiten für paarige Originalwerte und Zielhäufigkeiten

Es ist hierbei anzumerken, dass direkt eine symmetrische Verteilung für die Überlagerungsbeiträge gewählt wurde. Auf der x-Achse sind die anzuwendenden Überlagerungswerte von $-d$ bis d zu sehen. Auf der y-Achse sind die Originalwerte abgetragen. In der jeweiligen Zelle sind die einzelnen Übergangswahrscheinlichkeiten abzulesen. Für Originalwerte größer 7 ist die Zeile für Originalwerte gleich 7 maßgebend. Die Überlagerungsmatrix bildet den Kern der CKM. Ihre Anwendung wird im nächsten Abschnitt beschrieben.

3.2. Methodik und Verfahrensdurchführung

Die wichtigsten Bestandteile des Verfahrens werden in [Enderle, 2019] dargestellt. Ähnlich wie in [Wipke, 2018] beschrieben, lässt sich nun ein Algorithmus formulieren. Ausgangspunkt sind die in einer Statistik erhobenen Mikrodaten bzw. Mikrodatensätze. Das sind die plausibilisierten Einzeldatensätze.

1. Erzeugung der Originalwerte mit einem Auswertungstool

2. Cell-Key-Bestimmung aus Zufallszahlen innerhalb des Auswertungs-Tools

3. Lookup-Modul

a) Auslesen der Überlagerungswerte aus der Überlagerungsmatrix

b) Addieren der Überlagerungswerte und Originalwerte

Die folgende Abbildung 2 visualisiert den schehmatischen Ablauf des zuvor beschrieben Algorithmus. Im Wesentlichen werden zwei technische System benötigt, um das Verfahren zu realisieren. Zum einen wird ein Auswertungstool benötigt, welches die gespeicherten Mikrodaten in Tabellenform bringt. Mit Hilfe des Lookup-Moduls werden dann die beiden Eingangsgrößen - Originalwerte und Cell-Keys - verwendet, um die Überlagerungswerte aus der Überlagerungsmatrix zu bestimmen. Diese Überlagerungswerte werden letztlich auf die Originalwerte addiert und stellen damit die finalen Werte für die Veröffentlichung dar.

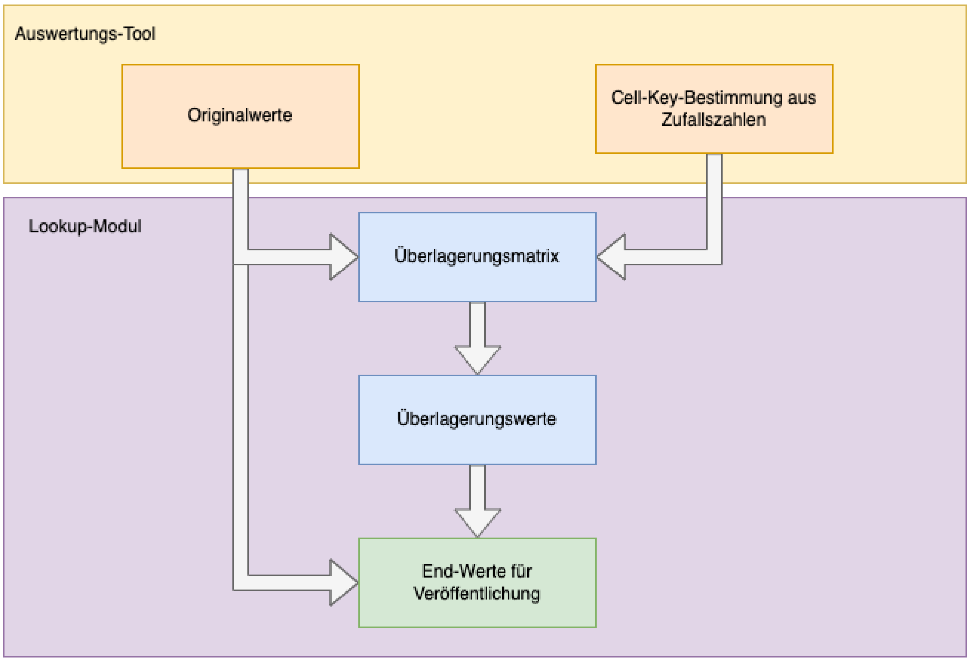

Abbildung 2: Ablaufdiagramm der Cell Key Methode

Die Einzelschritte des Verfahrens sollen nun im Detail beleuchtet werden.

3.2.1. Erzeugung der Originalwerte

Dieser Schritt ist spezifisch für die jeweilige Zieltablle, die veröffentlicht werden soll. Im Allgemeinen werden Filterungen anhand bestimmter Merkamle vorgenommen und dann Summen der Fallzahlen gebildet. Für diese Operation - also die Tabellierung - führt man die Beizeichnung f ein.

3.2.2. Cell-Key-Bestimmung

Für die Bestimmung des Cell-Keys wird jedem Mikrodatensatz zunächst eine gleichverteile Zufallszahl r, der sog. *Record-Key*, mit $r \sim \mathcal{U}(0,1)$ zugeordnet. Mit diesen Record-Keys wird dieselbe Auswertungstabelle wie mit den Originalwerten gebildet. Man führt also diesseble Operation f auf den Daten durch, wie in Abschnitt 3.1.1. Es ergeben sich also Summen von Record-Keys. Von diesen Record-Key Summen werden nur die Nachkommastellen betrachtet. Dieser Wert definiert den *Cell-Key* [Enderle, 2019]. Um dieses Vorgehen zu verdeutlichen soll das folgende Beispiel aus der Hochschulstatistik betrachtet werden. Tabelle 3 zeigt die Mikrodaten mit Record-Keys, wie sie in einer Datenbank gespeichert sein könnten.

ID	Universitätät	Geschlecht	Record-Key
1	Würzburg	m	0,611853
2	Eichstätt	w	0,139494
3	München	w	0,292145
4	München	m	0,366362
5	Würzburg	m	0,456070
6	Würzburg	m	0,785176
7	Bamberg	m	0,199674
8	München	w	0,514234
9	München	m	0,592415
10	München	m	0,046450

Tabelle 3: Mikrodaten mit Record-Keys

Tabelle 4 stellt die Daten nach Durchführung der Tabellierung f dar. Die Daten wurden hier nach der Universitätät und nach dem Geschlecht zusammengefasst. Die entsprechenden Fallzahlen, Record-Key-Summen und Cell-Keys werden abgebildet.

ID	Universitätät	Geschl.	Fallzahl	Record-Key-Summe	Cell-Key
1	Bamberg	m	1	0,199674	0,199674
2	Eichstätt	w	1	0,139494	0,139494
3	München	m	3	1,005227	0,005227
4	München	w	2	0,806379	0,806379
5	Würzburg	m	3	1,853099	0,853099

Tabelle 4: Aggregierte Daten mit Record-Key-Summen und Cell-Key

Die Werte in Zeile 5 von Tablle 4 ergeben sich durch Zählen der Zeilen in Tabelle 3, in denen $Universit\ddot{a}t = W\ddot{u}rzburg \wedge Geschlecht = m$. Die Record-Key-Summe in dieser Zeile ergibt sich nach $0{,}611853 + 0{,}456070 + 0{,}785176 = 1{,}853099$. Der daraus abgeleitete Cell-Key beträgt damit 0,853099.

3.2.3. Lookup-Modul

Im Anschluss an die Bestimmung der Originalwerte und der dazugehörigen Cell-Keys gilt es nun die Überlagerungen zu berechnen. Für die Bestimmung eines Überlagerungswertes dient das Paar ($Originalwert$, $CellKey$) als Input. Damit meint man den Wert einer einzelnen Tabellenzelle und den über dieselbe Operation f berechneteten Cell-Key. Das Lookup Modul stellt die Funktionalität bereit, anhand dieses Wertepaares den zugehörigen Überlagerungswert aus der Überlagerungsmatrix abzulesen. Um dies weiter zu illustrieren wird die Überlagerungsmatrix aus Abbildung 1 herangezogen. Betrachtet man Zeile 5 aus Tabelle 4, so liegt das Paar ($Originalwert = 3$, $CellKey = 0{,}853099$) vor. Um nun auch den dazu passenden Überlagerungswert zu bestimmen, wählt man die Zeile i der Überlagerungsmatrix, die dem $Originalwert$ - hier also 3 - entspricht. Anschließend bestimmt man mit Hilfe des $CellKey$ die Spalte j, für welche der Wert in der Matirx erstmalig kleiner als $CellKey$ ist. In diesem Fall führt dies zu $j = 1$, da $0{,}853099 < 0\ 86$. Der Wert $j = 1$ ist damit der zu addierende Überlagerungswert. Daraus ergibt sich ein finaler Wert von $3 + 1 = 4$. Wendet man dieses Vorgehen vollständig auf das Beispiel aus Tabelle 4 an, so ergeben sich die finalen und damit geheimgehaltenen Werte aus Tabelle 5.

ID	Universitätät	Geschl.	Fallzahl	Überlagerung	Finaler Wert
1	Bamberg	m	1	−1	0
2	Eichstätt	w	1	−1	0
3	München	m	3	−3	0
4	München	w	2	1	3
5	Würzburg	m	3	1	4

Tabelle 5: Überlagerungen und final geheimgehaltene Werte

In Anhang A ist eine *Python* Realisierung dieses Verfahrens zu finden.

3.3. Besonderheiten der Cell Key Methode

Eine Besonderheit, die sich aus der Verwendung der Cell Key Methode ergibt, ist die Nicht-Additivität des Verfahrens. Dadruch, dass Rand- und Zwischensummen nicht erst nach der Geheimhaltung, sondern ebenfalls während der Tabellierung gebildet werden, unterliegen auch diese Werte dem Geheimhaltungsmechanismus. Um dies zu veranschaulichen soll abschließend ein weiteres Fallbeispiel aus der Hochschulstatistik betrachtet werden. Tabelle 6 zeigt für eine Universitätät jeweils eine Insgesamt-Position i sowie die Unterteilung nach dem Geschlecht m und w. Es sind sowohl die Originalwerte als auch die final überlagerten Werte abgebildet. Auch die Positionen für die Zwischensummen i wurden mit der Tabellierungs-Operation f auf Basis der Originalwerte und der Record-Keys gebildet. Damit entsteht also ein eigner Cell-Key und Überlagerungswert für diese Zahlen. Für das Beispiel der Universität Bamberg aus Tabelle 6 sieht man schnell, dass für die Originalwerte die gewohnte Additivität vorhanden ist, denn $166 = 75 + 91$. Führt man diese Betrachtung allerdings auf den überlagerten Werten durch, so erhält man eine falsche Aussage, denn $168 \neq 75 + 91$. Diese Phänomen bezeichnet man als *Nicht-Additivität* einer statistischen Tabelle.

ID	**Universitätät**	**Geschl.**	**orig. Fallzahl**	**überl. Fallzahl**
1	Bamberg	i	166	168
2	Bamberg	m	75	75
3	Bamberg	w	91	91
4	Eichstätt	i	46	48
5	Eichstätt	m	17	20
6	Eichstätt	w	32	29

Tabelle 6: Beispiel zur Nicht-Additivität der CKM

Diese Nicht-Additivität wird beim Cell Key Verfahren in Kauf genommen, da im Weiteren zwei konkrete Vorteile dieser Methode überwiegen. Das unabhängige und separate Überlagern von Tabellenfeldern führt zu diesen wichtigen Vorteilen dieses Geheimhaltungsverfahrens [Enderle, 2019]. Zum einen wird eine *tabellenübergreifenede Konsistenz* erreicht. Die zu addierenden Überlagerungswerte zu einem bestimmten Originalwert (z.B. Studierende der Universitätät Bamberg) sind bei gleicher Datengrundlage immer identisch - unabhängig von der Zieltabelle. Dies ergibt sich aus den einmalig zugespielten Record-Keys und dem danach folgenden deterministeischen Lookup-Modul. Der zweite überwiegende Vorteil der CKM ist die *Genauigkeit* des Verfahrens [Enderle, 2019]. Dieser Punkt beschreibt die Vermeidung von zufällig gleichgerichteten Überlagerungen. Dies würde im Einzelfall zu größeren Veränderungen zwischen Original- und geheimgehaltenen Wert führen. Durch die Überlagerung aller Tabellenzellen umgeht man diese Gefahr.

3.4. Aufdeckungsrisiko

Damit die Cell Key Methode ein sicheres Geheimhaltungsverfahren bleibt, dürfen die Verfahrensparameter unter keinen Umständen der Öffentlichkeit preisgegeben werden. Sind beispielsweise die Maximalabweichung oder die Standardabweichung der Überlagerungen bekannt, so können in Tabellenfeldern, die einen additiven Zusammenhang aufweisen, Rückschlüsse auf die Originalwerte mit Hilfe der Randsummen gezogen werden. Für Details sei hier auf [Höhne, 2019] und [Giessing, 2016] verwiesen.

4. Fazit

Mit der Cell Key Methode wurde ein modernes Geheimhaltungsverfahren präsentiert, welches sich durch seine Flexibiltät - auch in der technischen Integration - von bestehenden Geheimhaltungsverfahren unterscheidet. So ist es ein hochatraktives Verfahren zur Gewährleistung der Geheimhaltung in neuen technischen System, wie der Auswertungsdatenbank der amtlichen Hochschulstatistik, oder auch dem anstehenden Zenus. Mit dem Wegfallen von manuellen Geheimhaltungsprozeduren lassen sich qualitativ hochwertige Ergenisse in deutlich kürzerer Zeit und mit übergreifenden Konsistenz veröffentlichen. Allerdings ist auch die Besonderheit der Nicht-Additivität deutlich an die Endnutzer der Daten zu kommunizieren, um eine Missinterpretation der Ergebnisse zu verhindern. Durch die wählbaren Verfahrensparameter, wird das CKM-Verfahren zukünftig sicher auch in anderen Teilen der amtlichen Statistik Einzug finden.

Literatur

Enderle, Tobias und Meike Vollmar: Geheimhaltung in der Hochschulstatistik. *WISTA* | 6, Statistisches Bundesamt (Destatis), Wiesbaden 2019.

Giessing, Sarah: Computational Issues in the Design of Transition Probabilities and Disclosure Risk Estimation for Additive Noise. *LNCS*, vol. 9867, Springer International Publishing, 2016.

Höhne, Jörg und Julia Höninger: Die Cell-Key-Methode ein Geheimhaltungsverfahren. *Statistische Monatshefte Niedersachsen* 1, 2019.

Nickl, Andreas: Datenschutz, Geheimhaltung, Anonymisierung. *Einfuührungsfortbildung* Bayerisches Landesamt für Statistik, Fürth, 2019.

Rothe, Patrick: Statistische Geheimhaltung Der Schutz vertraulicher Daten in der amtlichen Statistik - Teil 1: Rechtliche und methodische Grundlagen *Bayern in Zahlen* 5, Bayerisches Landesamt für Statistik, München, 2015.

Rothe, Patrick: Statistische Geheimhaltung Der Schutz vertraulicher Daten in der amtlichen Statistik - Teil 2: Herausforderungen und aktuelle Entwicklungen. *Bayern in Zahlen* 8, Bayerisches Landesamt für Statistik, München, 2015.

Wipke, Mirko: Geheimhaltung im Data Warehouse - Prototypische Implementierung von automatisierter Geheimhaltung im Data Warehouse für die amtliche Hochschulstatistik in Bayern. *Bayern in Zahlen* 12, Bayerisches Landesamt für Statistik, Fürth, 2018.

A. Python Implementierung

Nachfolgend ist die vollständige *Python* Implementierung des CKM Verfahrens für ein Testbeispiel aus der Hochschulstatistik abgebildet. Für die Werte der Überlagerungsmatrix wurde auf die Literatur zurückgegriffen.

```python
import math
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt


# Werte der  Überlagerungsmatrix und des Änderungsvektors stammen
# aus "Die Cell-Key-Methode  ein Geheimhaltungsverfahren"
# von Jörg Höhne und Julia Höninger.
OVERLAY_MATRIX = np.matrix([
    [0, 0, 0, 0, 0, 0, 0, 0, 0],
    [0, 0, 0, 0.6875, 0.6875, 0.6875, 0.9375, 1, 1],
    [0, 0, 0.3533, 0.3533, 0.3533, 0.9440, 0.9970, 0.9990, 1],
    [0, 0.1620, 0.1620, 0.1620, 0.6620, 0.8560, 0.9970, 0.9990, 1],
    [0.0870, 0.0870, 0.0870, 0.1920, 0.6920, 0.8590, 0.9970, 0.9990, 1],
    [0, 0, 0.1450, 0.3270, 0.8270, 0.8590, 0.8930, 0.9490, 1],
    [0, 0.0400, 0.1500, 0.2850, 0.7850, 0.8600, 0.9200, 0.9600, 1],
    [0.0200, 0.0600, 0.1450, 0.2500, 0.7500, 0.8550, 0.9400, 0.9800, 1]
])

CHANGE_VECTOR = [-4, -3, -2, -1, 0, 1, 2, 3, 4]


def matrix_plot(matrix, vector):
    """
    Heatmap der Überlagerungsmatrix.
    """
    sns.heatmap(matrix, annot=True, xticklabels=vector, cbar=False)
    plt.xlabel("Überlagerungen")
    plt.ylabel("Originalwerte")
    plt.show()


def generate_data(n, seed) -> pd.DataFrame:
    """
    Erstellt zufällige Testdaten von einigen Merkmalsaus-
    pärgungen. Jeder Eintrag wird mit einem gleichverteil-
```

```python
    ten Recordkey zwischen 0 und 1 versehen.
    """
    np.random.seed(seed)
    universities = [
        "Bamberg", "Wuerzburg",
        "Muenchen", "Eichstaett"]
    sex = ["m", "w"]

    uni_data = np.random.choice(universities, size=n, replace=True, p=[0.15, 0.3, 0.5, 0.05])
    sex_data = np.random.choice(sex, size=n, replace=True, p=[0.5, 0.5])
    record_key_data = np.random.uniform(low=0.0, high=1.0, size=n)

    return pd.DataFrame(
        list(zip(uni_data, sex_data, record_key_data)),
        columns =['university', 'sex', 'record_key']
    )


def tabulate_data(data, rollout=False):
    """
    Gibt die gruppierten Daten als Häufigkeitstablle mit den
    Summierten Recordkeys zurück.
    Wenn die `rollout` Option gesetzt ist, werden ebenfalls
    alle Gruppensummen mit ausgegeben.
    """
    grouped_data = data.groupby(["university", "sex"]).agg(["count", "sum"])
    grouped_data.columns = ["count", "record_key_sum"]
    grouped_data.reset_index(inplace=True)

    if rollout:
        rollout_data = data.loc[:, data.columns != "sex"].groupby(["university"]).agg(["count", "sum"])
        rollout_data.columns = ["count", "record_key_sum"]
        rollout_data.reset_index(inplace=True)
        rollout_data["sex"] = "i"
        rollout_data = rollout_data.iloc[:, [0,3,1,2]]

        sum_col = pd.DataFrame({
            "university": ["sum"],
            "sex": ["i"],
            "count": [grouped_data["count"].sum()],
```

```
            "record_key_sum": [grouped_data["record_key_sum"].sum()
    ]
        })

        grouped_data = grouped_data.append(
            [rollout_data, sum_col], ignore_index=True)
        grouped_data = grouped_data.sort_values(
            by=["university", "sex"])

    return grouped_data


def get_cell_key(value: float) -> float:
    """
    Gibt den Dezimalanteil ein Gleitkommazahl zurück.
    """
    return value - int(value)


def get_overlay_matrix_value(matrix, vector, values,
    record_key_sums, seed, p0=1) -> list:
    """
    Gibt eine Liste aus Überlagerungswerten passend zu den Paaren
    aus Originalwerten und Recordkey-Summen zurück. Aus den
    Recordkey-Summen werden die Cellkeys bestimmt.
    """
    np.random.seed(seed)
    overlay_col = []
    num_rows, _ = matrix.shape

    for value, record_key_sum in zip(values, record_key_sums):
        if value == 0:
            overlay_col.append(value)
            continue
        elif value < num_rows:
            cell_keys = matrix[value, :]
        else:
            cell_keys = matrix[num_rows - 1, :]

        for index, key in enumerate(cell_keys.tolist()[0]):
            if key > get_cell_key(record_key_sum):
                overlay_value = vector[index]
                break
        else:
            overlay_value = vector[-1]
```

```python

        if p0 is not None:
            overlay_value = np.random.choice([overlay_value, 0], size=1, p=[1 - p0, p0])[0]
        overlay_col.append(overlay_value)

    return overlay_col


def apply_ckm(data, matrix, vector, value_col_names, record_key_names, seed, p) -> pd.DataFrame:
    """
    Wendet die Cell Key Methode auf die angegebenen Spalten des
    Dataframes an. Hierfür wird der Überlagerungswert berechnet
    und zu den Spalten addiert.
    Gibt einen Dateframe mit den überlagerten Werten zurück.
    """
    output_data = data.copy()
    for col_name, record_key_name, p0 in zip(value_col_names, record_key_names, p):
        output_data[col_name] = data[col_name] + get_overlay_matrix_value(matrix, vector, data[col_name], data[record_key_name], seed, p0)
    return output_data


if __name__ == "__main__":
    data = generate_data(1001, 42)
    table_data = tabulate_data(data)
    overlayed_data = apply_ckm(table_data, OVERLAY_MATRIX, CHANGE_VECTOR, ["count"], ["record_key_sum"], seed=42, p=[0])

    # Erstellen der Hearmap.
    matrix_plot(OVERLAY_MATRIX, CHANGE_VECTOR)

    # Darstellen der Testdaten, der tabellierten Testdaten und der
    # geheimgehaltenen Testdaten.
    print(data)
    print(table_data)
    print(overlayed_data)
```

Listing 1: CKM Python Beispiel

Ich erkläre hiermit, dass ich die Seminararbeit mit dem Titel *Statistische Geheimhaltung: Cell Key Methode* im *Sommersemester 2022* selbständig angefertigt, keine anderen Hilfsmittel als die im Literaturverzeichnis genannten benutzt und alle aus den Quellen und der Literatur wörtlich oder sinngemäßübernommenen Stellen als solche gekennzeichnet habe.

Bamberg, den 25. Juli 2022
Unterschrift