Geheimhaltung im Data Warehouse

Prototypische Implementierung von automatisierter Geheimhaltung im Data Warehouse für die amtliche Hochschulstatistik in Bayern

Dipl. Kfm. Mirco Wipke

Die amtliche Hochschulstatistik befindet sich in einem umfangreichen Änderungsprozess. Mit der Novellierung des Hochschulstatistikgesetzes im Jahr 2016 wurden die bisherigen Statistiken erweitert und neue Statistiken eingeführt. Ferner wurde festgelegt, die Ergebnisse der Hochschulstatistiken in einer bundesweiten Auswertungsdatenbank vorzuhalten. Diese Auswertungsdatenbank wird derzeit entwickelt und wird nicht nur den Statistischen Ämtern dienen, sondern auch einem breiteren Nutzerkreis, weil eine (automatisierte) statistische Geheimhaltung implementiert wird. Das Bayerische Landesamt für Statistik nutzt bereits seit weit über einem Jahrzehnt ein Data Warehouse als Auswertungssystem für seine amtlichen Hochschulstatistiken. Vor zwei Jahren begann dort die prototypische Umsetzung eines automatisierbaren Geheimhaltungsverfahrens, für das nun Produktionsreife erreicht wurde.

Vorbemerkung

Das Bayerische Landesamt für Statistik (LfStat) nutzt bereits seit 2003 ein Data Warehouse (DWH) als Auswertungssystem, um die Daten der amtlichen Hochschulstatistiken flexibel nach unterschiedlichen Merkmalen bzw. Merkmalskombinationen und Zeitständen oder Zeitverläufen zu tabellieren oder in Grafiken ausgeben zu können. Inzwischen ist auch eine Funktion für die automatische Umsetzung statistischer Geheimhaltung integriert, bei der die wesentlichen Nutzungseigenschaften des DWH wie performante, flexible und inhaltlich valide Auswertungen erhalten bleiben.

Diese Integration erfolgte im Zusammenhang mit der Entscheidung der Statistischen Ämter des Bundes und der Länder, künftig ein einheitliches Geheimhaltungsverfahren in der Hochschulstatistik anzuwenden. Daher wird im ersten Abschnitt dieses Artikels das Auswahlverfahren kurz erläutert und das gewählte Verfahren beschrieben. Dadurch erklärt sich, weshalb die spezifischen fachlichen Anforderungen mit dieser Methode am besten umgesetzt werden können. Im zweiten Abschnitt wird die Implementierung eines solchen Verfahrens mit beispielhaft gewählten Parametern im DWH beschrieben. Dabei lässt sich die Wirkungsweise des Geheimhaltungsverfahrens instruktiv mit den ohnehin in einem DWH genutzten Aufbauschritten der Datenmodellierung und -beladung verbinden.

Der Beschreibung liegen Erfahrungen des LfStat aus einer prototypischen Implementierung zugrunde, bei der das dargestellte Geheimhaltungsverfahren mit beispielhaft gewählten Parametern erfolgreich getestet wurde. Derzeit werden im Statistischen Verbund Qualitätsberechnungen durchgeführt, auf deren Grundlage die später in der Auswertungsdatenbank Hochschulstatistik zu implementierenden Geheimhaltungsparameter bestimmt werden sollen. Die spezifischen Parameter werden intern durch die Statistischen Ämter festgelegt und nicht veröffentlicht. In der gewählten Umsetzungsvariante ist das dargestellte Geheimhaltungsverfahren somit technisch einsatzbereit.

Entscheidungsfindung für ein DWH-Geheimhaltungsverfahren

Rechtsgrundlagen

Rechtsgrundlage für die amtliche Hochschulstatistik ist das Hochschulstatistikgesetz (HStatG)[1], das seit seiner Novellierung im Jahr 2016 in § 8 HStatG auch den Aufbau einer bundesweiten Auswertungsdaten-

1 Hochschulstatistikgesetz vom 2. November 1990 (BGBl. I S. 2414), das zuletzt durch Artikel 3 des Gesetzes vom 7. Dezember 2016 (BGBl. I S. 2826) geändert worden ist.

bank vorsieht. Diese Datenbank wird nicht nur den Statistischen Ämtern, sondern auch einem breiteren Nutzerkreis zur Verfügung stehen, da eine (automatisierte) statistische Geheimhaltung umgesetzt wird.

Die Forderung nach Geheimhaltung entspringt § 16 Bundesstatistikgesetz (BStatG). Demnach sind Einzelangaben über persönliche und sachliche Verhältnisse, die für eine Bundesstatistik angegeben werden, von den jeweils durchführenden statistischen Stellen geheim zu halten, soweit nichts anderes bestimmt ist.

Dieses Statistikgeheimnis ist für die amtliche Statistik verpflichtend und ein Pendant zur Auskunftspflicht nach § 15 BStatG. Denn die Auskunftspflicht, ausgestaltet in den jeweiligen speziellen Statistikgesetzen, verursacht einen staatlichen Eingriff in die Freiheitsrechte natürlicher oder juristischer Personen (beispielsweise in die informationelle Selbstbestimmung). Im Gegenzug sind die mittels Eingriffsverwaltung erhaltenen Informationen bestmöglich zu schützen, d. h. sie sind in einer Form zu anonymisieren, die keine Rückschlüsse mehr auf Betroffene erlaubt. Damit einher geht ein Informationsverlust, den es bestmöglich zu begrenzen gilt.

Geheimhaltungsmethodische Erwägungen: Gegenüberstellung dreier Verfahren

Die Aufgabenstellung lautete, ein Geheimhaltungsverfahren zu finden, das die Anforderungen im Spannungsfeld aus Datenschutz und Datennutzbarkeit hinreichend berücksichtigt. Daher wurden folgende Prüfpunkte formuliert:

1. Schutz: Aufdeckungsrisiko minimieren
2. Ergebnisqualität: Geringe Datenveränderung, wenig Informationsverlust
3. Implementierung: Wirtschaftlichkeit des Umsetzungsaufwands
4. Auswertung: Hohe Flexibilität – ohne Einschränkungen und koordinierte Absprachen
5. Praktikabilität: Dezentrale und unabhängige Anwendbarkeit
6. Nutzerakzeptanz: Gute Vermittelbarkeit
7. Einheitlichkeit: Standardisierte Geheimhaltung in Bund und Ländern

Im Fokus stehen dabei die nicht-monetären Hochschulstatistiken, die persönliche und studiumsrelevante Daten zu den Studierenden wie beispielsweise Alter, Studienfach sowie abgelegte Prüfungen umfassen. Ferner auch Daten zum Hochschulpersonal, zu Habilitierten und neuerdings auch zu Promovierenden. Geheimhaltungsrelevant sind hier hauptsächlich Fallzahlen.

Diese Fallzahlen werden beispielsweise in den statistischen Berichten bisher überwiegend mit Zellsperrverfahren geheim gehalten. Tabellenzellen mit weniger als drei Ausprägungen werden durch einen Punkt gesperrt (Primärsperrung). Um eine Aufdeckung der Primärsperrung zu vermeiden, sind meist weitere Zellen zu punkten (Sekundärsperrung). Die Zellsperrung ist zwar als Verfahren bekannt und leicht verständlich, jedoch fehlt die benötigte Automatisierbarkeit für den Einsatz in einem DWH mit flexibler Auswertung. Alternativ wurden daher drei datenändernde Verfahren verglichen, nämlich deterministisches Runden, eine pre-tabulare Datenänderung und eine post-tabulare stochastische Überlagerung.

Deterministisches Runden zur Basis 5 ist in der Variante des kaufmännischen Rundens allgemein bekannt. Knapp zusammengefasst besagen bereits die Namen der beiden anderen Verfahren, dass post-tabulare Methoden Tabellen nach deren Erstellung überlagern, während pre-tabulare Vorgehensweisen die Daten bereits ändern, bevor sie tabelliert werden. In Tabelle 1 sind die drei genannten Verfahren gegenübergestellt und anhand der zu Beginn dieses Abschnitts aufgeführten Prüfungsaspekte in der Beschreibung gegliedert. Der Aspekt „Einheitlichkeit“ entfällt, weil er mit allen drei Verfahren gleichermaßen erreichbar ist.

Da die fachlich gewünschten Eigenschaften nicht im erforderlichen Ausmaß abgedeckt werden, wird der Detailvergleich ohne die pre-tabulare Änderung fortgesetzt.

Deterministisch oder stochastisch – Detailvergleich und Entscheidung

Die post-tabulare stochastische Überlagerung wird in einer Ausgestaltung näher betrachtet, die durch das Australian Bureau of Statistics[2] entwickelt und daher oft mit „ABS“ verschlagwortet wurde. Die offizielle

2 Fraser B., Wooton J. (2006): A proposed method for confidentialising tabular output to protect against differencing, WP. 35, Joint UNECE/ Eurostat work session on statistical data confidentiality, Geneva 2005.

Bezeichnung lautet im Statistischen Verbund inzwischen „Stochastische Überlagerung mittels Cell-Key-Methode[3]“ (CKM). Die Idee der CKM ist, eine bereits erstellte Auswertung systematisch so zu modifizieren, dass insbesondere kleine Fallzahlen relativ stärker verändert werden als große Fallzahlen, um so Geheimhaltungsprobleme zu lösen, die besonders im Zusammenhang mit flexibler Tabellierung entstehen, wie es ein DWH erlaubt.

Zur Entscheidungsfindung wurden das deterministische Runden und CKM mit unterschiedlichen Parametrisierungen gerechnet. Der CKM-Parameter „Maximalabweichung“ ist die absolute Datenänderung, die maximal in einer Tabellenzelle auftreten kann; die weiteren Parameter „Varianz“ und „Bleibewahrscheinlichkeit“ seien hier zunächst nur genannt. Sie werden wegen der besseren Anschaulichkeit später im Zusammenhang mit der sogenannten Übergangsmatrix behandelt.

Das deterministische Runden wurde neben dem bekannten Runden zur Basis 5 auch mit Varianten zur Basis 3, 7, 9 sowie 10 parametrisiert.

Die auf diese Weise unterschiedlich parametrisierten Verfahren bzw. Varianten lassen sich anhand folgender Gütekriterien beurteilen:

- Mittelwert der absoluten Abweichungen: Durchschnitt der tatsächlichen Änderungen
- Empirische Bleibewahrscheinlichkeit: Anteil der unveränderten Tabellenzellen
- Informationsverlust: Kennzahl, die auf Basis der Hellinger-Distanz[4] zwischen Originalverteilung sowie überlagerter Verteilung gebildet wird
- Tatsächliche Aufdeckung: Algorithmisch ermittelter Anteil der aufgedeckten Felder

Diese Gütekriterien wurden in einem aufwändigen, mehrstufigen Verfahren anhand von Daten des Zensus 2011 gerechnet und erbrachten zusammengefasst folgende Ergebnisse: Alle Rundungsverfahren hinterlassen im Vergleich zu den betrachteten CKM-Varianten einen höheren Informationsverlust, da wenige Werte unverändert bleiben. Trotzdem besteht, zumindest beim 3er- und 5er-Runden, ein relativ hohes Aufdeckungsrisiko. Die CKM schneidet in diesem Vergleich durch geringen Informationsverlust und geringstmögliches Aufdeckungsrisiko deutlich besser ab. Auf der

3 Der Begriff „cell key“ wird im folgenden Abschnitt bei der Darstellung der CKM erläutert.

4 Die Hellinger-Distanz quantifiziert die Ähnlichkeit zwischen zwei Verteilungen: je geringer die Distanz, desto ähnlicher die Verteilungen.

Tab. 1 **Übersicht zu drei Geheimhaltungsverfahren**

Beschreibungsaspekte	Deterministisches Runden	Post-tabulare Überlagerung*	Pre-tabulare Änderung*
1. Schutz (Primär und Sekundär)			
Aufdeckungsrisiko (des Veränderungsmusters)	relativ groß	gering	kein Risiko
2. Ergebnisqualität			
Konsistenz (tabellenübergreifend)	ja	ja	ja
Additivität	nein	nein	ja
Informationsverlust	Rundung aller Werte	sehr gering	gering
3. Implementierung(saufwand)	mittel bis gering	initial sehr hoch	initial hoch
4. Auswertung(sflexibilität)			
Vorabfestlegung von Auswertungen nötig?	nein	nein	ja
5. Praktikabilität			
Müssen Daten zentral verändert werden?	nein	nein	ja
Können Daten dezentral ausgewertet werden?	ja	ja	ja
6. Vermittelbarkeit/Nutzerakzeptanz	einfach	anspruchsvoll	anspruchsvoll

* Der Darstellung der post-tabularen Überlagerung liegt die Cell-Key-Methode, der pre-tabularen Änderung das SAFE-Verfahren (vgl. Gießing et al. (2014)) zugrunde.

Quelle: Statistisches Bundesamt und eigene Beschreibung,
vgl. auch: Rohde J., Seifert C., Gießing S. (2018): Entscheidungskriterien für die Auswahl eines Geheimhaltungsverfahrends, WISTA – Wirtschaft und Statistik, Ausgabe 03/2018.

Antal L., Enderle T., Giessing S. (2017): Harmonised protection of census data in the ESS, unter: ec.europa.eu/eurostat/cros/system/files/methods_for_protecting_census_data.pdf, zuletzt abgerufen am 17.12.2018.

Gießing S., Heinzl F., Kleber B., Wilke A. (2014): Geheimhaltung beim Zensus 2011, WISTA – Wirtschaft und Statistik, Ausgabe 11/2014.

anderen Seite hat das deterministische Runden den großen Vorteil, dass es allgemein bekannt und somit vergleichsweise wenig erklärungsbedürftig ist, während CKM methodisch und technisch deutlich aufwändiger ist und insofern auch anspruchsvoller in der Vermittlung. Nachdem CKM jedoch geheimhaltungsmethodisch überlegen ist, stellt es die beste Option da.

Im Jahr 2018 entschieden sich die fachlich für die Hochschulstatistik zuständigen Referentinnen und Referenten der Statistischen Ämter des Bundes und der Länder abschließend für den Einsatz der CKM und erteilten den Auftrag, die Verfahrensparameter auf die Hochschulstatistik zu optimieren.

Die Implementierung der mit dem Hochschulstatistikgesetz angeordneten zentralen Auswertungsdatenbank, in die eine CKM integriert werden soll, ist in der Vorbereitungsphase. Die prototypische Umsetzung der CKM in das bayerische Hochschulstatistik-DWH demonstriert die technische Realisierbarkeit eines solchen Geheimhaltungsverfahrens in einer marktüblichen Business-Intelligence-Software.

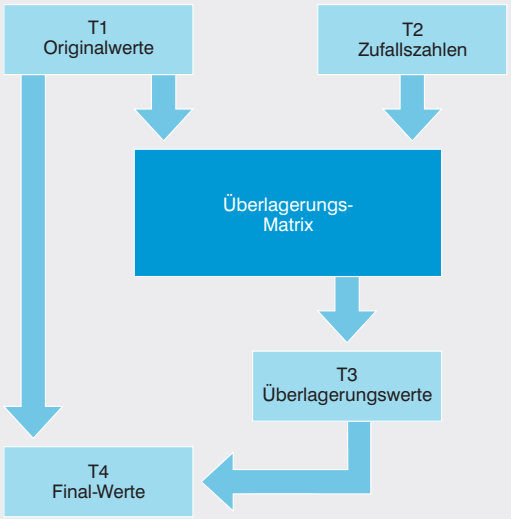
Die CKM: Erläuterung anhand der Implementierung

Überblick

Die CKM gehört – wie bereits erwähnt – zur Gruppe der post-tabularen stochastischen Überlagerungen und ist „post-tabular“, weil erst eine (Original-)Auswertung erstellt und danach eine Überlagerung angewandt wird. „Stochastisch“ wird die Überlagerung dadurch, dass die Eingangsdaten in jedem Datensatz einmalig um eine gleichverteilte Zufallszahl ergänzt werden, die später in die Überlagerung einfließt. Dieser „seed“ oder „record key“ genannte Wert zwischen 0 und 1 bewirkt, dass das Verfahren im ersten Schritt stochastisch ist, weil sich damit eine zufällige Überlagerung erzielen lässt. Im zweiten Schritt erreicht es Konsistenz, weil deterministisch stets derselbe dem Datensatz zugeordnete Zufallswert in die Auswertung eingeht.

Insofern lässt sich die CKM hinsichtlich Ergebniserstellung in vier Schritte gliedern, wobei zur Anschaulichkeit hier zunächst von Tabellen gesprochen wird. Der CKM-Flow-Chart in Abbildung 1 zeigt Ablauf und Zusammenhänge:

Abb. 1
CKM-Flow-Chart



1. **Originalwert-Tabelle erzeugen:** Erstellen einer Tabelle mit Originalwerten (z. B. Studierende im aktuellen Wintersemester). Vereinfachte Annahme: Jeder Datensatz, also jede Zeile des Auswertungsmaterials, wird mit 0 oder 1 gezählt, je nachdem, ob der Datensatz berücksichtigt werden soll[5]. Damit entsteht der Tabellenwert aus einer Aufsummierung über diesen Datensatzzähler.
2. **Zufallszahl-Tabelle erzeugen:** Erstellen derselben Auswertung wie mit den Originalwerten, aber nun auf Basis der Zufallszahlen: Statt des oben genannten Datensatzzählers wird die gleichverteilte Zufallszahl (zwischen 0 und 1) aufsummiert.
3. **Überlagerungswert-Tabelle erzeugen:** Die in ihrem Aufbau identischen Tabellen aus den Schritten 1 und 2 gehen als Input in eine sogenannte Überlagerungsmatrix ein. Als Output entsteht eine wieder strukturgleiche Tabelle von Überlagerungswerten. Jede Tabellenzelle gibt an, wie der Originalwert geändert wird.
4. **Final-Tabelle: Originalwert-Tabelle mittels Überlagerungswert-Tabelle modifizieren:** Die Überlagerungswert-Tabelle wird auf die Originalwert-Tabelle angewandt.

Um die Auswirkungen in den Endergebnissen zu verdeutlichen, enthält Tabelle 2 zwei CKM-Reportbeispiele aus dem bayerischen Hochschulstatistik-DWH.

5 Dieser Mechanismus wird im nachfolgenden Abschnitt detailliert erklärt, soll aber hier nicht den Blick auf die Idee des Verfahrens verstellen.

Tab. 2 **CKM-Reportbeispiele zu tabellenübergreifender Konsistenz und Nicht-Additivität**

Beurlaubte Studierende in Bayern im Wintersemester 2016/17 nach ausgewählten Hochschularten

a) insgesamt

Hochschularten	Beurlaubte Studierende CKM
Universitäten	8 936
Fachhochschulen	1451
Gesamt	10 387

b) nach Geschlecht

Hochschularten	Beurlaubte Studierende CKM			**Vergleichs-summierung der Innenfelder**
	m	w	Gesamt	
Universitäten	3 486	5 448	8 936	**8 934**
Fachhochschulen	556	893	1 451	**1 449**
Gesamt	4 042	6 345	10 387	
Vergleichssummierung der Innenfelder	**4 042**	**6 341**	**10 383**	

Ersichtlich sind zwei CKM-Charakteristika:

- **Tabellenübergreifende Konsistenz:** entsteht, weil identische Inhalt stets identisch überlagert (oder beibehalten) werden: Die Zeilensummen in Tabelle 2b entsprechen den Tabelleninnenfeldern von Tabelle 2a. In beiden Tabellen werden z. B. für Universitäten 8 936 Beurlaubte ausgewiesen.
- **Nicht-Additivität:** entsteht, weil Summen (bzw. Randfelder) eigens bzw. separat überlagert werden: Die Zeilensummen in Tabelle 2b entsprechen nicht der Summe der Tabelleninnenfelder, z. B. teilen sich an Universitäten 8 936 Beurlaubte in 3 486 männliche und 5 448 weibliche Fälle, was zusammen 8 934 und nicht 8 936 Beurlaubte ergibt.

Angemerkt sei, dass beurlaubte Studierende eher selten analysiert werden. Aufgrund ihrer recht geringen Fallzahlen sind sie jedoch als Beispiel dankbar, weil so Abweichungen recht leicht zu erkennen sind. Im bayerischen Hochschulstatistik-DWH wurden sie als erste Überlagerungs-Implementierung gewählt, um Performance-Beeinträchtigungen zunächst außer Acht lassen zu können. Die inzwischen realisierten CKM-Studierenden-Kennzahlen lassen indessen vermuten, dass die automatische Geheimhaltung nicht als Verzögerung in der Ergebnisbereitstellung am Bildschirm wahrnehmbar wird.

Studierendenstatistik und Warehousing

Nach dem grundlegenden Einstieg auf der Grob-Ebene folgt nun ein Gang durch die Details. Tabelle 3 zeigt einen Ausschnitt aus dem Datensatz der Studierendenstatistik. Jede Zeile im Datenmaterial repräsentiert einen Studierenden (Kopf-Statistik), jede Spalte ein Merkmal, dessen Ausprägungen (alpha-)numerisch codiert sind.

Die Daten werden in dieser Form in das DWH importiert[6]. Im Zuge des weiteren Ladeprozesses werden aus den Importdaten „Fakten“ erzeugt (vgl. Infokasten auf Seite 848), die als Grundlage für Kennzahlenberechnungen dienen. Tabelle 4 veranschaulicht diesen Schritt und enthält denselben Datensatz wie Tabelle 3, ergänzt um Fakten-Felder bzw. -Spalten. In Tabelle 5 werden aus diesen Fakten Kennzahlen bzw. „Metriken“ (vgl. Infokasten auf Seite 850) gebildet, z. B. als Summe.

6 Zusätzlich werden Schlüsselverzeichnisse aufgenommen, um aus den Codierungen Klartexte zu erzeugen, was jedoch für die Geheimhaltung nicht von Belang ist.

Tab. 3 **Ausschnitt aus dem Datensatz der Studierendenstatistik**

EF21_AnzahlHochschulsemester	EF28_EinschreibungsArt	EF32_AnzahlFachsemester	EF33_angestrebterAbschluss	EF35_Stgg1_RSZ	EF36_Stgg1_SF1
06	3	06	182	6	021
01	1	01	182	6	021
07	3	01	182	6	021
08	4	07	788	4	021

Bedeutung der Signierungen:

Art der Einschreibung
1: Ersteinschreibung
3: Rückmeldung
4: Beurlaubung

Angestrebter Abschluss
182: Erststudium Bachelor
788: Konsekutiver Master

Regelstudienzeit (RSZ)

1. Studienfach (SF) im 1. Studiengang (Stgg) 021: BWL

Fakten schaffen bedeutet in einem DWH, die Merkmale eines Datensatzes auf bestimmte Inhalte zu verdichten. Das einfachste Beispiel ist, das Vorliegen einer Eigenschaft, z. B. „studierend“, mit 0 als „Nein“ und 1 als „Ja“ zu erfassen.

Es handelt sich um eine etablierte Form der DWH-Datenmodellierung, und anders als in diesem Artikel der Anschaulichkeit halber beschrieben, werden normalerweise nicht die Ursprungsdaten um Fakten erweitert, sondern Fakten sind Tabellen, in die die Ursprungsdaten „übersetzt“ und geladen werden.

Das Übersetzen wird im Artikel u. a. anhand der Studierenden-Eigenschaft beschrieben, die sich aus dem Merkmal „Einschreibungsart“ ableitet: „WENN EF28_EinschreibungsArt <4, DANN f_Studierende=1“. Dass die zeitaufwändigen Wenn-Dann-Auswertungen bereits beim Laden erfolgen, bewirkt performant laufende Auswertungen im Live-System.

Dimensionen stellen die Organisationsform für Attribute wie z. B. „Hochschulart“, „Hochschule“ oder „Studienfach“ dar und ermöglichen die Differenzierung nach Merkmalen samt Klartextangaben. Wie bei den Fakten handelt es sich letztlich um Datenbank-Tabellen, die über Schlüsselbeziehungen mit den Fakten verbunden sind, und über die der Nutzer im DWH die Metriken differenzieren kann. Dimensionen sind für die hier dargestellte Variante der CKM-Implementierung irrelevant, weil kein technischer Anpassungsprozess nötig ist.

Tab. 4 **Datensatzausschnitt der Studierendenstatistik, ergänzt um Fakten-Felder**

EF21_AnzahlHochschulsemester	EF28_EinschreibungsArt	EF32_AnzahlFachsemester	EF33_angestrebterAbschluss	EF35_Stgg1_RSZ	EF36_Stgg1_SF1		f_Studierende	f_StudAnf_HS1	f_StudAnf_FS1	f_Studierende: WENN EF28_EinschreibungsArt <4, DANN f_Studierende=1 f_StudAnf_HS1: WENN (EF28_EinschreibungsArt =1 UND EF21_AnzahlHochschulsemester = 01), DANN f_StudAnf_HS1=1 f_StudAnf_FS1: WENN (EF28_EinschreibungsArt <4 UND EF32_AnzahlFachsemester = 01), DANN f_StudAnf_FS1=1
06	3	06	182	6	021		1	0	0	
01	1	01	182	6	021		1	1	1	
07	3	01	182	6	021		1	0	1	
08	4	07	788	4	021		0	0	0	

Tab. 5 **Beispiele für Metriken (Originärwerte)**

Metrik	Berechnungsvorschrift	Ergebnis
Studierende	Summe (f_Studierende)	3
Studienanfänger im 1. Hochschulsemester	Summe (f_StudAnf_HS1)	1
Studienanfänger im 1. Fachsemester	Summe (f_StudAnf_FS1)	2

Anwendungslogik der CKM

Nachdem CKM im Flow-Chart der Abbildung 1 zum besseren Überblick in vier Schritten umrissen wurde, wird die Darstellung unter Anwendung der Fakt-Metrik-Logik nun in Tabelle 6 zu sieben Schritten erweitert.

In Tabelle 7 wird das Datensatzbeispiel wiederum fortgesetzt. Es ist dort der erste CKM-Schritt integriert, nämlich indem bei jedem Datensatz der außerhalb des DWH erzeugte „seed“ in der Spalte „Zufallszahl“ ergänzt wird. Derzeit ist dies ein individueller Baustein für das bayerische DWH. Sobald das Verfahren auch verbundweit eingesetzt wird, wird die Zufallszahl durch die Fachanwendung, mit der die Einzeldaten nach Übermittlung durch die Auskunftspflichtigen an die amtliche Statistik plausibilisiert und aufbereitet werden, erzeugt und bereitgestellt.

In der DWH-Logik bedeutet die CKM-Implementierung, dass Originär-Fakten-Felder um Zufalls-Fakten („record keys“) erweitert werden. Tabelle 7 enthält daher zusätzlich zur Spalte „Zufallszahl“ auch drei Zufalls-Fakten. Statt einer 1 als Datensatzzähler wird dort die Zufallszahl des Datensatzes geschrieben. Analog zu den Originär-Fakten werden die Zufallszahlen bzw. Zufallszahl-Fakten zur Erzeugung von Zufallszahl-Metriken (zu „cell keys“) aufaddiert, nur dass hier eben Dezimalbrüche summiert werden[7]. Angewandt auf das Beispiel spiegelt Tabelle 8 die Metrikergebnisse.

Das vierzeilige Datensatzbeispiel hat hiermit seinen didaktischen Zweck erfüllt, und für die weitere Erörterung wird zu einer Echtauswertung gewechselt: Tabelle 9 zeigt durch Nebeneinanderstellen

7 Die vereinfachte Darstellung, dass jede Datensatzzeile der Lieferdaten auch im DWH einer Datensatzzeile entspricht, vernachlässigt Vor-Aggregationen zur Performance-Verbesserung. Geheimhaltungsmethodisch ist dies unerheblich, da Zufallszahl- und Originär-Fakten gleich behandelt werden, weshalb sich Zwischen-Aggregationen nicht auf das Verfahren auswirken.

Tab. 6 **Umsetzungsschritte der CKM**

Umsetzungsschritt	Pendant zum Beispiel in Abbildung 1
1. Erweiterung jedes Datensatzes um eine Zufallszahl („seed“/ „record key“)	
2. Bildung einer Originär-Metrik	T1 Originalwerte
3. Bildung einer Zufallszahl-Metrik („cell key“) analog zur Originär-Metrik (mit Zufallszahl-Fakten)	T2 Zufallszahlen
4. Normierung der Zufallszahl-Metrik auf das Intervall [0;1]	
5. Einspeisen der normierten Zufallsmetrik in die Überlagerungsmatrix	Überlagerungsmatrix
6. Erzeugen des Überlagerungswerts (Output der Überlagerungsmatrix)	T3 Überlagerungswerte
7. Überlagerung der Originär-Metrik	T4 Final-Werte

Tab. 7 **Datensatzausschnitt, ergänzt um Zufallszahl und Zufallszahl-Metriken**

EF21_AnzahlHochschulsemester	EF28_EinschreibungsArt	EF32_AnzahlFachsemester	EF33_angestrebterAbschluss	EF35_Stgg1_RSZ	EF36_Stgg1_SF1		f_Studierende	f_StudAnf_HS1	f_StudAnf_FS1		Zufallszahl	fZZ_Studierende	fZZ_StudAnf_HS1	fZZ_StudAnf_FS1
06	3	06	182	6	021		1	0	0		0,63	0,63	0	0
01	1	01	182	6	021		1	1	1		0,45	0,45	0,45	0,45
07	3	01	182	6	021		1	0	1		0,16	0,16	0	0,16
08	4	07	788	4	021		0	0	0		0,87	0	0	0

Metriken sind Messzahlen oder Kennzahlen. Metrik-Funktionen wie Summe, Durchschnitt, Minimum, Maximum etc. werden auf Fakten angewendet.

Um das obige Beispiel fortzusetzen: Die Summe über den Fakt „Studierende“ ergibt die Gesamtzahl der Studierenden.

Der Gesamtprozess besteht aus einem datenbankseitigen Laden der Fakten und einem nutzerorientierten Anbieten von zumeist vordefinierten Metriken im Front-End und schafft in beiden Bereichen „Laufzeitvorteile“, indem individuelle „Auswertungsprogrammierungen“ minimiert (oder ganz unterbunden) werden. Dieser hohe Grad an Automatisierung und Standardisierung trägt dadurch zu inhaltlicher Stabilität und somit durchaus auch zu Validität bei.

von Metriken die Umsetzungsschritte aus Tabelle 6 im bayerischen Hochschulstatistik-DWH. Schritt 1 wurde vorangehend erörtert, ist Teil des DWH-Ladevorgangs, zum Zeitpunkt der Tabellenerzeugung abgeschlossen und daher hier nicht mehr darstellungsrelevant. Folgerichtig beginnt Tabelle 9 mit Schritt 2 und zeigt die erzeugte Originär-Metrik. In Schritt 3 folgt die Zufallszahl-Metrik, welche in Schritt 4 normiert wird, was hier bedeutet, dass die Vorkommastellen verworfen werden. Die Ermittlung des Überlagerungswertes (Schritte 5 und 6) werden im nachfolgenden Abschnitt behandelt.

Für Endnutzer wird das DWH letztlich, wie bisher, nur eine Metrik ausgeben, allerdings dann eben eine überlagerte wie in Tabelle 2 bereits praktiziert. Die anderen Metriken werden im Echtbetrieb technisch im Hintergrund gebildet.

Die Übergangsmatrix und ihre Umsetzung im DWH

Das Kernstück der CKM ist die Übergangsmatrix, die in diesem Artikel bisher als Blackbox behandelt wurde. Beim nun folgenden Blick in die Blackbox stellt (und beantwortet) sich die Frage, wie eine Mechanik beschaffen sein muss, damit aus den Input-Faktoren „Originär-Metrik“ und „Zufallszahl-Metrik“ als Output ein Überlagerungswert entsteht. Dies unter Berücksichtigung folgender Design-Eigenschaften:

a) Kleine Fallzahlen sollen gezielt häufiger und – relativ zu ihrem (kleinen) Originalwert stärker verändert werden.
b) Das stochastische Element soll die Geheimhaltung durch Erzeugen von Unsicherheit unterstützen.
c) Die Ergebnisse sollen erwartungstreu sein, d. h. bei gegebener Originalhäufigkeit i soll der Erwartungswert der Abweichung zwischen Originalwert und überlagerter Häufigkeit Null sein.

Typischerweise erfolgen Input/Output-Betrachtungen anhand von Tabellen. Die Kopfzeile und die Vorspalte definieren den Input, während die Tabellen-Innenfelder den Output ergeben. Bei der CKM hingegen greift die Vorspalte i den Input aus der Originär-Metrik auf und die Tabellen-Innenfelder den Input aus der Zufallszahl-Metrik. Der Output „Überlagerungswert“ ergibt sich aus der Differenz von Kopfzeile j und Vorspalte i. Diese komprimierte Beschreibung wird nachfolgend in Einzelschritte aufgelöst und durch ein Beispiel veranschaulicht. Dabei wird einerseits die Erzeugung der Matrix behandelt, andererseits ihre konkrete Nutzung.

Das Lesen der Übergangsmatrix – ob durch Mensch oder Maschine – beginnt in der Vorspalte i. Hier wird eine Fallunterscheidung getroffen: Werte (einer Originär-Metrik) von 0 bis 3 werden differenziert, während

Tab. 8 **Beispiele für Metriken (Originärwerte und Zufallszahlen)**

Originärmetrik: Summe von …		**Zufallszahlmetrik: Summe von …**	
f_Studierende	3	fZZ_Studierende	1,24
f_StudAnf_HS1	1	fZZ_StudAnf_HS1	0,45
f_StudAnf_FS1	2	fZZ_StudAnf_FS1	0,61

Tab. 9 **Report-Ausschnitt mit CKM-Metriken**

Schritt ... aus Tabelle 6	2	3	4	6	7
Hochschule	Beurlaubte Studierende	Zufallszahl Beurlaubte Studierende	Normierte Zufallszahl Beurlaubte Studierende	Überlagerungswert Beurlaubte Studierende	CKM Beurlaubte Studierende
U Bamberg	767	382,458546	0,458546	0	767
U Bayreuth	577	285,425074	0,425074	0	577
U Passau	860	443,041765	0,041765	-2	858
Kath. U Eichstätt-Ingolstadt	90	44,293874	0,293874	0	90
Augustana-H Neuendettelsau (ev)	7	4,177469	0,177469	0	7
U der Bundeswehr München	24	11,397436	0,397436	0	24
U Erlangen-Nürnberg	1.263	622,904768	0,904768	0	1.263
U München	1.673	858,963365	0,963365	2	1.675
U Würzburg	800	385,976906	0,976906	3	803
U Regensburg	740	374,500657	0,500657	0	740
U Augsburg	622	312,923576	0,923576	0	622
TU München	1.514	750,522354	0,522354	0	1.514
H für Politik München	6	1,917059	0,917059	0	6
H für Philosophie München (rk)	7	4,265035	0,265035	0	7
H für Musik Würzburg	36	16,643603	0,643603	0	36
H für Musik und Theater München	61	31,771829	0,771829	0	61
Akademie der Bildenden Künste München	35	16,987209	0,987209	3	38
Akademie der Bildenden Künste Nürnberg	8	4,114842	0,114842	0	8
H Fresenius Idstein (Priv. FH)	66	33,665617	0,665617	0	66
FH Augsburg	74	37,688904	0,688904	0	74
FH Coburg	9	3,758837	0,758837	0	9
FH München	316	156,533725	0,533725	0	316
Techn. H Nürnberg Georg Simon Ohm (FH)	346	191,852741	0,852741	0	346

für Werte von 4 oder größer keine Unterscheidung mehr getroffen wird. Dies entspricht der Design-Eigenschaft a. Tabelle 10 zeigt eine Beispielmatrix, in der zusätzlich die Forderung berücksichtigt wurde, dass eine Null stets eine Null bleibt. Tatsächlich Gezähltes kann also bis zum Maximalwert geändert werden, nicht Vorhandenes wird (durch diese Parametrisierung der CKM) jedoch nicht künstlich erschaffen.

Der nächste Schritt des Ablesens läuft über die Matrix-Innenfelder. Sie repräsentieren pro Zeile eine Zufallsverteilung, also eine Funktion zwischen 0 und 1. Es handelt sich dabei um die Eintrittswahrscheinlichkeiten der Überlagerungswerte. Die in Tabelle 9 dargestellte Normierung der Zufallszahl-Metrik – bisher nur als notwendiger Anwendungsschritt beschrieben – wird nun in ihrer Funktionsweise deutlich. Eine Zufallszahl-Metrik größer 1 muss normiert werden, damit diese im Wertebereich der Verteilung (in der Matrix-Zeile) liegt, und zusammen mit der „Einstiegskoordinate“ i die Verortung von j definiert. Dadurch ist gleichzeitig die Designeigenschaft b, die Wirkung des stochastischen Elements, beschrieben.

Aus Tabelle 10 lässt sich allerdings der Überlagerungswert nicht unmittelbar ablesen. An dieser Stelle

Tab. 10 **Beispielhafte Übergangsmatrix**

i\j	0	1	2	3	4	5	6
0	1	0	0	0	0	0	0
1	0,51	0	0,46	0,03	0	0	0
2	0,17	0	0,55	0,24	0,04	0	0
3	0	0	0,42	0,28	0,18	0,12	0
4	0	0	0,0739	0,2442	0,3637	0,2442	0,0739

Quelle: Statistisches Bundesamt im Rahmen des Projekts „OPEN SOURCE TOOLS FOR PERTURBATIVE CONFIDENTIALITY METHODS“, Specific grant agreement n° 2018.0108, ec.europa.eu/eurostat/cros/content/perturbative-confidentiality-methods_en, zuletzt abgerufen am 17.12.2018.

gabelt sich die Darstellung, und die angewandte Praxis wird kurz zurückgestellt, um die Darstellung der Matrixerzeugung abzuschließen. Für die Erzeugung der Übergangsmatrix sind Parameter wie die Maximalabweichung (D), die Bleibewahrscheinlichkeit (p) für einen Originär-Tabellenwert oder die Varianz, also die Streuung der Verteilung (einer Matrix-Zeile), vorzugeben. Die konkreten Ausprägungen der Matrix werden dann, unter der Nebenbedingung „Erwartungstreue“, über eine nicht-lineare Optimierung ermittelt. Die Designeigenschaft c wird also durch den Optimierungsalgorithmus berücksichtigt. Damit ist die Logik der Matrixerzeugung für das Verständnis des Gesamtzusammenhangs hinreichend beschrieben, und die Darstellung widmet sich nun wieder dem Ablesen des Überlagerungswertes.

Tabelle 11 zeigt die (zeilenweise von links nach rechts) kumulierten Einzelwerte der Übergangswahrscheinlichkeiten. Die Zelleinträge stellen die Obergrenze des Intervalls dar, in die die jeweilige Ausprägung der normierten Zufallsmetrik fällt. Die jeweils vorausgehende Zelle markiert die Untergrenze. Das Ablesen des Überlagerungswertes aus der Übergangsmatrix in Tabelle 11 lässt sich zusammen mit Tabelle 9 am Beispiel der Universität Passau demonstrieren:

- Die Originär-Metrik „Beurlaubte Studierende“=860 ergibt i=4.
- Die normierte Zufallszahl-Metrik=0,0417… ist kleiner als der Zellenwert 0,07 und größer als der Zellenwert 0 in Zeile 4 und führt damit zu Spaltenindex j=2.
- Konsequenz:
  - Der Überlagerungswert ergibt sich durch $(j-i) = (2-4) = -2$.
  - Die überlagerte Originär-Metrik lautet 858 $(=860-2)$.

Abschließend sei kurz eine Umsetzungsvariante umrissen: Wird die Übergangsmatrix in eine Überlagerungsmatrix wie in Tabelle 12 überführt, lässt sich per Datenbankfunktion darauf zugreifen. Es gibt zwei Übergabevariablen: Input-Parameter …

- Originär-Metrik (wie z. B. „Beurlaubte Studierende“) definiert i,
- normierte Zufallszahl-Metrik (z. B. „Zufallszahl Beurlaubte Studierende“) definiert cum_zuf,

… sodass der Überlagerungswert (changeValue) als Rückgabewert ausgelesen werden kann. Hiermit ist nun auch der oben zunächst zurückgestellte Schritt 5 aus Tabelle 6 (Umsetzungsschritte der CKM) beschrieben.

Resümee

Die amtliche Hochschulstatistik setzt zu einem großen Sprung nach vorne an: Es wird eine Auswertungsda-

Tab. 11 **Übergangsmatrix (aus Tabelle 10) mit kumulierten Übergangswahrscheinlichkeiten**

i\j	0	1	2	3	4	5	6
0	1	0	0	0	0	0	0
1	0,51	-	0,97	1	-	-	-
2	0,17	-	0,72	0,96	1	-	-
3	0	0	0,42	0,7	0,88	1	-
4	0,00	0,00	0,07	0,32	0,68	0,93	1,00

Tab. 12 **Ausschnitt aus einer datenbanktauglichen Überlagerungsmatrix**

i	cum_zuf	changeValue
0	1	0
1	0,51	-1
1	0,97	1
1	1	2
2	0,17	-2
2	0,72	0
2	0,96	1
2	1	2
3	0,07	-2
3	0,31	-1
...		

tenbank geben, die umfangreiche Analysen u. a. über Studierende, Promovierende, Prüfungen und Personal erlauben wird. Dieses Auswertungssystem wird einem breiten Publikum zugänglich sein, da dort ein automatisierbares Geheimhaltungsverfahren zum Einsatz kommen wird. Gleichzeitig wird damit eine bundesweit einheitliche Geheimhaltung umgesetzt.

Der Statistische Verbund hat mit Empfehlung des Ausschusses für die Hochschulstatistik[8] als einheitliches Geheimhaltungsverfahren eine post-tabulare stochastische Überlagerung in der Variante der sogenannten CKM beschlossen. Dieses in der deutschen amtlichen Statistik noch neue Verfahren verändert kleinere Fallzahlen stärker als höhere Fallzahlen. Während das bekannte Zellsperrverfahren kleine Fallzahlen ganz ausblendet und für einen konsistenten Schutz weitere Zellinhalte eliminieren muss, ist es gerade die Stärke der CKM, keine Sperrungen sowie auch insgesamt wenig Informationsverlust zu verursachen.

Aufgrund der separaten Überlagerung von Tabellenzellen und deren Summen ist CKM nicht additiv. Insbesondere bei kleinen Häufigkeitswerten springt diese Nicht-Additivität ins Auge. Es verlangt insofern durchaus einen Kulturwandel, diese geringfügigen Differenzen je Tabellenzelle bis zur Höhe der Maximalabweichung zu akzeptieren. Zwar ist die Belastbarkeit kleiner Fallzahlen generell kritisch einzustufen, auch ohne datenändernde Geheimhaltung. Allerdings ist dies auch kontextabhängig: Abweichungen bei Studierenden in Ägyptologie werden von den meisten Datennutzern sicher anders gewertet als ein Rückgang von Professorinnen in Naturwissenschaften.

CKM verursacht bei Menschen, die das Verfahren zum ersten Mal kennenlernen, durchaus Fragen: Konterkariert ein „Verrauschen“ von amtlichen Zahlen nicht den Aufwand, der in die Aufbereitung und Plausibilisierung investiert wird? Schadet es nicht der Reputation, wenn eine amtliche Statistik nicht exakt ist, sodass Summen nicht „aufgehen“? Die Fragen lassen sich getrost verneinen. Letztlich geht es im Grunde um ein altes Thema, das bereits beim Zellsperrverfahren hinsichtlich der Bewertung, was veröffentlichungssensible Merkmale sind, diskutiert wurde: Guter Datenschutz bei gleichzeitig guter Datenqualität ist (und bleibt) ein Spannungsfeld.

Folgerichtig ist die amtliche Statistik auch damit befasst, die Vermittlung der CKM vorzubereiten. Dazu gehört u. a. die Frage, wie viel von diesem Geheimhaltungsverfahren, insbesondere dessen Parametrisierung, veröffentlicht werden darf, um es einerseits Datennutzern verständlich zu erklären, ohne es andererseits in der Geheimhaltungskraft zu schwächen.

Die (prototypische) Implementierung im bayerischen Hochschulstatistik-DWH zeigt jedenfalls, dass das Geheimhaltungsverfahren CKM in eine marktverfügbare Business Intelligence Software technisch integriert und damit auch in der Praxis – unter vernachlässigbarem Aufwand für Eigenprogrammierung – umgesetzt werden kann.

Die Automatisierung und Vereinheitlichung der Geheimhaltung wird in den Statistischen Ämtern voraussichtlich zu einer wesentlichen Erleichterung in der Bearbeitung von Sonderauswertungen führen, und bereits eine allgemeine Verfügbarkeit der Auswertungsdatenbank wird wohl die Zahl der Individualbeantwortungen senken. Im Gegenzug dürfte die Unterstützung von Datennutzern bei der Bedienung der Auswertungsdatenbank wohl (Mehr-)Aufwand bewirken. Dass das Verfahren auch dezentral angewandt werden kann, verschafft den Statistischen Landesämtern Planungssicherheit in der eigenen Organisation und stärkt gleichzeitig ein einheitliches, methodisches Fundament.

8 Vgl. § 12 I HStatG.