Geheimhaltung

Das Geheimhaltungsverfahren SAFE

von **Jörg Höhne**

Der vorliegende Beitrag ist eine Methodenbeschreibung des Anonymisierungsverfahrens SAFE. Mit dem Verfahren SAFE kann ein anonymer Datenbestand erzeugt werden. Das kann einerseits mit dem Ziel erfolgen, einen anonymisierten Einzeldatenbestand über die Forschungsdatenzentren herauszugeben, beispielsweise als sogenanntes Scientific-Use-File. Andererseits kann SAFE als pre-tabulares Geheimhaltungsverfahren eingesetzt werden. Anstatt nach der Tabellenerzeugung alle Angaben in Tabellenfeldern zu prüfen und einzelne zu sperren, werden bei pre-tabularen Geheimhaltungsverfahren alle Tabellen aus dem anonymen Datenbestand berechnet und schützen so die Einzelangaben der Befragten. Der Beitrag beschreibt den mathematischen Hintergrund und die Lösungsalgorithmen.

1. Einleitung

Die Statistischen Ämter des Bundes und der Länder erheben Daten aus allen Bereichen des gesellschaftlichen Lebens – für über 250 Bundes- und Landesstatistiken, aber auch für Wahlen und Volksabstimmungen. Sie bereiten diese Daten auf und werten sie aus. Die Einzelangaben von Personen, wirtschaftlichen Einheiten und anderen Merkmalsträgern werden von der amtlichen Statistik geschützt und bleiben den Nutzerinnen und Nutzern der Statistiken daher verborgen. Das soll einen Missbrauch der Einzelangaben verhindern und so die Bereitschaft zur wahrheitsgemäßen Auskunft erhalten. In den Statistikgesetzen, z. B. § 16 des Bundesstatistikgesetzes[1], ist dieses Vorgehen gesetzlich geregelt.

Traditionell gewährleisten Geheimhaltungsverfahren den Schutz der Einzelangaben durch Informationsreduktion. Um die Geheimhaltung zu realisieren, werden in statistischen Ergebnissen (post-tabular) die Einzelangaben einerseits nach bestimmten Kriterien zusammengefasst (z. B. Wirtschaftszweige, Betriebsgrößenklassen, Altersgruppen, Regionen) und andererseits Angaben in noch vorhandenen sensiblen Tabellenfeldern unterdrückt oder durch weitere Vergröberung von Gliederungen unsichtbar gemacht. Außerdem besteht die Möglichkeit, durch Vergröberung der Angaben (Rundung) den Nutzen der Tabellenfelder zur Aufdeckung von Einzelangaben bei einem Missbrauch zu reduzieren.

Ein anderer Weg der statistischen Geheimhaltung besteht darin, sicherzustellen, dass bereits die Einzeldaten nicht mehr ihren Merkmalsträgern zugeordnet werden können. Die Geheimhaltungsverfahren werden dabei bereits vor der Auswertung/Tabellierung angewandt (pre-tabulare Verfahren). Das erfolgt beispielsweise durch das Entfernen von Informationen, die für die Reidentifikation besonders kritisch sind oder durch das gezielte Verändern einzelner Merkmale. Das Verfahren SAFE ist ein pre-tabulares Verfahren, bei dem eine anonyme Version des Datenkörpers über Datenveränderung der Einzelangaben erstellt wird. Aus diesem Datenkörper können dann alle potenziellen Auswertungen erstellt werden. In keiner Auswertung tritt dann mehr ein Geheimhaltungsfall auf. Da alle Auswertungen aus derselben anonymen Quelle erfolgen, sind sie außerdem untereinander konsistent.

a | Klassifikation der Geheimhaltungsverfahren

	Informationsreduzierende Verfahren	Datenverändernde Verfahren
Pre-tabulare Verfahren	• Vergröberung (Zusammenfassen von Kategorien) • Entfernen von Merkmalen	• Mikroaggregation, z. B. SAFE • Swapping • Stochastische Überlagerung auf Mikrodatenebene
Post-tabulare Verfahren	• Zellsperrung • Zusammenfassung	• Deterministische (konventionelle) Rundung • Zufällige Rundung • Kontrollierte Rundung • Stochastische Überlagerung auf Tabellenfeldebene
	↓ Löschen oder unterdrücken Information (auch unkritische Felder bei Sekundärsperrungen)	↓ Schutz entsteht durch Unsicherheit (auch bei unkritischen Feldern)

1 Gesetz über die Statistik für Bundeszwecke (Bundesstatistikgesetz – BStatG) vom 22. Januar 1987 (BGBl. I S. 462, 565), zuletzt geändert durch Artikel 13 des Gesetzes vom 25. Juli 2013 (BGBl. I S. 2 749).

Jedes Geheimhaltungsverfahren wird daran gemessen, dass es einen ausreichenden Schutz der Einzelangaben bietet und dabei mit einem möglichst geringen Informationsverlust einhergeht. Dass es bei Verfahren der Zellsperrung einen Informationsverlust gibt, ist für die Nutzerinnen und Nutzer von Tabellen leicht ersichtlich, da unterdrückte Felder keine Information mehr enthalten. Der Schutz der Einzelangaben durch datenverändernde Verfahren beruht auf der Änderung von Werten, die die Verwertbarkeit beim Missbrauch einschränkt. Die Datenveränderungen sind aber auch in Auswertungen enthalten und stellen dort ebenfalls einen Informationsverlust dar. Der Umfang dieses Informationsverlustes kann durch Qualitätsmaße angegeben werden. Er wird im Beitrag an mehreren Stellen angesprochen.

SAFE ist ein Verfahren der Mikroaggregation. Bei Mikroaggregationsverfahren werden einzelne sich unterscheidende Datensätze einer Mikrodatendatei durch gezielte Auswahl und Gruppenbildung so vereinheitlicht, dass jeder Datensatz in der Basisdatei mit mindestens zwei weiteren Sätzen in der Datei identisch ist. Die hier beschriebene Version des Verfahrens ist zur Behandlung von kategorialen Merkmalen geeignet. Der vorliegende Beitrag ist die überarbeitete Version von Höhne (2003), der die aktuellen Weiterentwicklungen bei der Anonymisierung von kategorialen Merkmalen einschließt. Das in der Programmversion von 2003 enthaltene SAFE-Modul zur Anonymisierung quantitativer Merkmale ist in der aktuellen Version nicht mehr enthalten. Das liegt darin begründet, dass im Rahmen vergleichender Untersuchungen Mikroaggregationsverfahren bei stetigen Werten im direkten Vergleich zu anderen datenverändernden Verfahren nicht gleich gut überzeugten, sodass die Entwicklungsaktivitäten in diesem Bereich auf andere Verfahren konzentriert wurden (siehe z. B. Ronning et al. 2005).[2] Der Beginn der Arbeiten zum SAFE-Verfahren liegt in den frühen 1990er-Jahren (vgl. Appel et al. 1993). Der Name entstand als Akronym für Sichere Anonymisierung für Einzeldaten (SAFE).

Bei Mikroaggregationsverfahren werden die zu vereinheitlichenden Gruppen meist durch die Minimierung eines Abstandsmaßes zwischen den Einheiten gebildet.[3] Die diversen Mikroaggregationsverfahren unterscheiden sich dabei in der Wahl des Abstandsmaßes, in der Gruppengröße und in der Art, wie die Gruppen nach der Gruppenbildung vereinheitlicht werden. Das Verfahren SAFE zählt zu den Mikroaggregationsverfahren, da auch hier Gruppen vereinheitlicht werden. Allerdings werden die Gruppen durch ein numerisches Optimierungsverfahren so gebildet, dass ein Satz an vorgegebenen Auswertungstabellen möglichst exakt – jedoch ohne Geheimhaltungsfälle – aus dem anonymen Material wieder erzeugt werden kann.

Das dargestellte Verfahren erfüllt das Kriterium der k-Anonymität mit k = 3, es ist somit 3-anonym. Ein Datenbestand ist k-anonym, wenn jede Merkmalskombination mindestens als k-Tupel auftritt. Originale Beobachtungseinheiten können Datensätzen des k-anonymen Datenbestandes nicht mehr eindeutig zugeordnet werden, da mindesten k-1 Datensätze genauso wahrscheinlich zum Original passen (vgl. Sweeney 2002).

2. Begriffsbestimmungen

a) Die Mikrodatendatei

Eine Mikrodatendatei ist eine Datei, in der jedes statistische Objekt (Merkmalsträger) durch einen einzelnen Datensatz (Zeile) repräsentiert wird. Sie wird deshalb auch als Einzeldaten- oder Basisdatei bezeichnet. Sie bildet den Ausgangspunkt für alle möglichen Auswertungen des Datenbestandes. Die Merkmale in Datensätzen unterscheiden sich hinsichtlich ihres Skalenniveaus, es gibt qualitative (kategoriale) Merkmale und quantitative (stetige) Merkmale. Da die aktuelle Version von SAFE nur kategoriale Merkmale anonymisiert, sollte die Mikrodatendatei nur die kategorialen Merkmale des zu anonymisierenden Datenbestandes enthalten. Die übrigen Merkmale der Mikrodatendatei (Identifikationsmerkmale oder quantitative Merkmale) werden durch das Verfahren nicht behandelt. Für Analysezwecke können sie in der Datei enthalten bleiben, um beispielsweise die Datenänderungen im Verfahren zu quantifizieren. Außerdem kann durch das nachträgliche Anwenden von Anonymisierungsverfahren für quantitative Merkmale eine Mikrodatendatei mit anonymen qualitativen und quantitativen Merkmalen erzeugt werden.

Bei qualitativen Merkmalen, auch kategoriale Merkmale genannt, handelt es sich um Merkmale, die eine diskrete, feste Anzahl an Ausprägungen haben. Die möglichen Ausprägungen sind in einer Schlüsseltabelle zusammengefasst. Handelt es sich um hierarchische Schlüssel, so können qualitative Merkmale durch Umschlüsselung auf höhere Aggregationsebenen umgesetzt werden, beispielsweise können Regionalschlüssel, wie Gemeinde, auch in Kreis, Regierungsbezirk oder Land umgeschlüsselt werden. Analog kann der Wirtschaftszweig in Branchen oder die einzelne Nationalität in Deutsch/Nichtdeutsch umgeschlüsselt werden.

Quantitative Merkmale sind in der amtlichen Statistik, bedingt durch die Messgenauigkeit, meist ganzzahlig. Es gibt keine endliche vorher festgelegte Schlüsselmenge, die zulässig ist. Beispiele sind Umsatz und Beschäftigte. Aus quantitativen Merkmalen lassen sich durch Gruppierung wieder qualitative Klassen erzeugen (z. B. Betriebe mit unter 20 Beschäftigten, 20 bis unter 50 Beschäftigten usw.). Mit der aktuellen Version von SAFE können keine quantitativen Merkmale anonymisiert werden.

Identifikationsmerkmale (Ident-Nummern, Betriebsnummern, Adress-IDs usw.) sind Schlüsselmerkmale, die eine eindeutige Zuordnung des

2 Im Bereich der Anonymisierung von wirtschaftsstatistischen Daten lag der Schwerpunkt der methodischen Forschung des Autors und seiner Kolleginnen im Amt für Statistik Berlin-Brandenburg bei der „kontrollierten stochastischen Überlagerung".

3 Eine Ausnahme bilden die Verfahren der stochastischen Mikroaggregation (siehe Lechner und Pohlmeier 2003).

Datensatzes zum statistischen Objekt ermöglichen. Diese werden im Rahmen der Anonymisierung nicht betrachtet, da sie vor der Nutzung anonymer Daten in jedem Fall entfernt werden.

Eine Mikrodatendatei (Basisdatei) B ist eine Menge aus n Objekten. Jedes Objekt wird durch ein Tupel von k Werten beschrieben. Dabei beschreiben die Werte $b_{i1}, \ldots, b_{ik}$ die qualitativen (kategorialen) Merkmale des i-ten Objektes.

B – Basisdatei mit $B = \{b_{ij}\}$ - $i = 1,2,\ldots,n$
- $j = 1,2,\ldots,k$

n – Anzahl der statistischen Objekte
k – Anzahl der qualitativen/kategorialen Merkmale
b_{ij} – Wert des Merkmals j beim Objekt i

b) Die Kontrolltabellen

Das Verfahren SAFE optimiert bei der Anonymisierung die Lösung an einem vorgegebenen Kanon an Auswertungstabellen. Auswertungstabellen sind dabei Häufigkeitstabellen, die dadurch gebildet werden, dass die Basisdatei über eine bestimmte Merkmalskombination aggregiert wird. Es erfolgt eine Gruppenbildung (Aufsummierung von Sätzen mit gleicher Ausprägungskombination). Diese geplanten Auswertungen werden Kontrolltabellen genannt, da die Qualität der Anonymisierung am möglichst guten Erhalt dieser Auswertungen kontrolliert wird.

Die Kontrolltabellen können einerseits automatisch erzeugt werden (alle möglichen Tabellen bis zur Dimension d). Dabei werden alle Tabellen aus Kreuzkombinationen der qualitativen Merkmale mit beliebigen Schlüsselstufen bis zu einer festgelegten Tabellendimension (Anzahl zusammen tabellierter Merkmale) verstanden. Bei einer Festlegung auf die maximale Dimension $d = 3$ werden beispielweise alle möglichen eindimensionalen Häufigkeitstabellen, alle zweidimensionalen Tabellierungen und alle möglichen Kreuztabellierungen aus drei qualitativen Merkmalen als Kontrolltabellen erstellt.

Alternativ zur automatischen Generierung besteht andererseits die Möglichkeit, Kontrolltabellen per Liste dem Programm zuzusteuern. Beide Varianten können auch kombiniert werden. Dadurch besteht auch die Möglichkeit, ausgewählte höherdimensionale (z. B. vier- und fünfdimensionale) Tabellen zusätzlich zu allen ein- bis dreidimensionalen Tabellen zu kontrollieren. Da aufgrund der hohen Anzahl der theoretisch möglichen höherdimensionalen Tabellen es meist nicht möglich ist, z. B. alle vier- und fünfdimensionalen Tabellen als Kontrolltabellen zu behandeln, ist diese Mischung der Kontrolltabellenfestlegung erforderlich, wenn auch einzelne hochdimensionale Tabellen später ausgewertet werden sollen.

3. Enthüllungsrisiken und der Lösungsansatz von SAFE

a) Enthüllungsrisiken in Tabellen

Aus § 16 BstatG ergibt sich die Verpflichtung, „Einzelangaben über persönliche und sachliche Verhältnisse […] geheimzuhalten". Jede Tabelle muss von der amtlichen Statistik daher vor der Veröffentlichung dahin gehend geprüft werden, dass kein Tabellenfeld (auch Tabellenzelle genannt) dazu geeignet ist, auf Einzelangaben zurück zu schließen. Von einem Enthüllungsrisiko oder einem Geheimhaltungsproblem spricht man, wenn aus einem Tabellenfeld Rückschlüsse auf ein einzelnes statistisches Objekt (Unternehmen, Bürger usw.) gezogen werden könnten und so Informationen über das statistische Objekt nur aufgrund der statistischen Veröffentlichung zugänglich werden.

Folgende Enthüllungsrisiken können bei der Veröffentlichung von Tabellen entstehen (vgl. Hundepool et al. 2012):

- Fallzahlprobleme
  Enthüllungsrisiken durch zu kleine Fallzahlen treten vor allem bei Wertetabellen auf. Wenn nur ein oder zwei Merkmalsträger zu einem Tabellenwert, beispielsweise einer Summe, beitragen, besteht das Risiko einer exakten Enthüllung von Einzelwerten. Trägt nur ein Merkmalsträger zu einem Tabellenfeld bei, so entspricht der Tabellenwert der Einzelangabe des Merkmalsträgers. Bei zwei statistischen Objekten besteht das Risiko darin, dass eines der beiden Objekte durch Differenzbildung die Information über das andere Objekt (aufgrund der gleichen Merkmale in der Regel der Konkurrent) problemlos generieren kann.
  Bei Häufigkeitstabellen zeigen kleine Fallzahlen seltene oder einzigartige Merkmalskombinationen an. Direkte Rückschlüsse auf Einzelangaben sind im Allgemeinen nicht möglich. Die Information, dass die Merkmalskombination selten oder einzigartig ist, kann eventuell trotzdem als problematisch eingestuft werden:
  – Die Identifikation könnte mithilfe von Zusatzwissen möglich sein.
  – Durch die Verknüpfung mit Informationen aus anderen Datenquellen/anderen Tabellen wird es möglich, direkt Rückschlüsse auf andere Merkmale der betreffenden Individuen zu ziehen.
- Randwerte/Randsummenprobleme
  Randsummenprobleme entstehen, wenn innerhalb einer Tabelle in einer Zeile oder Spalte nur eine Zelle belegt ist. In diesem Fall können, auch wenn mehr als zwei Objekte zur konkreten Ausprägung des Tabellenwertes beitragen, Attribute für alle Beitragenden enthüllt werden. Ein Beispiel in der Todesursachenstatistik für einen Randwert und ein daraus entstehendes Randsummenproblem ist das Folgende: Innerhalb einer Region und Altersgruppe sterben alle Personen an der gleichen Krankheit. Allein die Information über das Alter und die Region einer gestorbenen Person ermöglicht es dann, die Todesursache anhand der Statistik eindeutig zuzuordnen. Randsummenprobleme sind immer inhaltlich zu betrachten, d. h. es ist zu entscheiden, ob das Merkmal geheimhaltungskritisch ist. Es ist

offensichtlich kein Geheimhaltungsproblem, wenn innerhalb der Gruppe der unter 6-Jährigen einer Region alle Kinder nicht erwerbstätig sind.

Neben den Problemen aus der Darstellung bei Tabellen existieren bei der Herausgabe von Mikrodaten weitere Reidentifikationsprobleme, die sich aus sogenannten „Matching"-Versuchen ergeben. Bei diesen wird versucht, Informationen, die man über statistische Objekte aus externen Quellen gewonnen hat, gegen Sätze der anonymen Basisdatei anzuspielen und so bei einer eineindeutigen Übereinstimmung zusätzliche Eigenschaften aus der Basisdatei abzulesen. Es wird üblicherweise unterschieden zwischen:

- Einzelangriffen – Datenangriffe, bei denen versucht wird, durch ein Matching für einzelne Objekte (z. B. Unternehmen) Informationen zu erhalten – und
- Massenfischzügen – hier wird ein Datenbestand an einen anderen gematcht und so möglichst viele Sätze zugeordnet (vgl. Lenz 2010, Ronning et al. 2005, Höhne 2010).

Allen vier Deanonymisierungsrisiken wird im Rahmen des SAFE-Verfahrens Rechnung getragen.

b) Lösungsansatz von SAFE

Das Verfahren SAFE erzeugt einen anonymen Datenbestand, der das Kriterium der k-Anonymität (vgl. Sweeney 2002) mit $k = 3$ erfüllt. Jede Ausprägungskombination tritt mindestens als Dreier-Tupel auf. Damit ergeben sich für die einzelnen Deanonymisierungsrisiken folgende Sicherheiten:

- Fallzahlprobleme können in den Tabellen nicht mehr auftreten, da mindestens drei Sätze zu einem Tabellenwert beitragen. Das bedeutet, dass entweder die in der Realität auftretenden kritischen Merkmalskombinationen entfernt wurden oder durch die Aggregation die Häufigkeit der Ausprägungskombination auf mindestens 3 erhöht wurde.
- Matching-Algorithmen, egal ob Einzelangriffe oder Massenfischzüge, können nur zu einer mehrdeutigen Zuordnung führen. Wenn ein Satz mehrere Entsprechungen in der anonymisierten Basisdatei hat, die wiederum durch Gruppenbildung entstanden sind, so kann auch nicht daraus geschlossen werden, dass die zusätzlich ablesbaren Eigenschaften für das Original gelten. Die k-Anonymität (vgl. Sweeney 2002) schützt vor Matching, da jede Ausprägungskombination dreifach vorhanden ist.
- Randsummenprobleme können bei Auswertungstabellen entstehen, aber auch hier ist es durchaus möglich, dass diese Probleme nur das Ergebnis der Anonymisierungstechnik sind. Künstliche Randsummenprobleme werden zusätzlich erzeugt, wenn durch das Gruppieren Objekte mit qualitativ verschiedenen Eigenschaften, aber geringen Häufigkeiten aus dem Datenbestand entfernt wurden. Es ist somit kein sicherer Rückschluss auf die Eigenschaften der Basisdatei mehr möglich.

Mit der oben eingeführten Notation lässt sich das allgemeine SAFE-Geheimhaltungsproblem („allgemein" bedeutet auch für Datenbestände mit stetigen Merkmalen) folgendermaßen darstellen:

Die originale Mikrodatendatei sei die Matrix B^o. Für diese Datei lässt sich für alle vereinbarten Schlüsselstufen und Aggregationsvorschriften die Menge aller vorgegebenen Auswertungstabellen bilden.

$T^o = F(B^o)$ – Matrix der Ergebnisse aller Auswertungstabellen.

$t^o_{p,q,m} = f_{p,q,m}(B^o)$ – Für jeden Tabellenwert $t^o_{p,q,m}$ des Merkmals m für die q-te Ausprägungskombination in der Auswertungstabelle p existiert eine Berechnungsfunktion $f^o_{p,q,m}$, mit der sich der Tabellenwert aus der Matrix der Mikrodaten bestimmen lässt. Übliche Funktionen sind die Summenfunktion zur Bildung von Aggregaten, aber auch Funktionen zur Berechnung von Durchschnitten oder Anteilswerten. Nach Bestimmung aller Geheimhaltungsfälle in den Auswertungstabellen lassen sich für die Geheimhaltungsfälle eine untere Grenze ($z^u_{p,q,m}$) und eine obere Grenze ($z^o_{p,q,m}$) bestimmen, die ein Unzulässigkeitsintervall um den geheim zu haltenden Wert beschreiben. Handelt es sich bei den Auswertungstabellen um Wertetabellen, so können die Unzulässigkeitsintervalle beispielsweise in Anlehnung an die Dominanzregeln für Wertetabellen abgeleitet werden, die unterstellen, dass ein beitragender Einzelwert nicht genauer als mit einem Fehler von x % rückschließbar sein darf. Handelt es sich bei den Auswertungstabellen um Fallzahltabellen, so darf keine eindeutige Zuordnung möglich sein. Damit sind für diese Tabellen die Fallzahlen 1 und 2 unzulässig.

Für eine anonymisierte Basisdatei B^a muss gelten:

$$T^a = F(B^a) \quad (1)$$

mit

$$t^a_{p,q,m} = f_{p,q,m}(B^a)$$

und

$$z^u_{p,q,m} \geq f_{p,q,m}(B^a) \quad \vee \quad z^o_{p,q,m} \leq f_{p,q,m}(B^a)$$

mit:

T^a – Matrix der Ergebnisse aller anonymen Auswertungstabellen.

B^a – Matrix der anonymisierten Basisdatei. Die anonymisierte Basisdatei ist dadurch gekennzeichnet, dass jede Zeile mindestens dreimal identisch in der Matrix enthalten ist.

$t^a_{p,q,m} = f_{p,q,m}(B^a)$ – In der Auswertungstabelle p wird für die Auswertung der Ausprägungskombination q beim Merkmal m der anonyme Tabellenwert $t^a_{p,q,m}$ bei Auswertung der anonymen Basisdatei über diese Funktion ermittelt.

$z^u_{p,q,m}, z^o_{p,q,m}$ – untere und obere Schranke des Unzulässigkeitsintervalls im Tabellenfeld. Wenn ein Geheimhaltungsfall beim Merkmal m in der Ausprägungskombination q der Tabelle p existiert, dann beschreiben diese Schranken Grenzen, ab denen der veröffentlichte Wert nicht mehr für einen Datenmissbrauch als nutzbar angesehen werden kann. Wenn kein Geheimhaltungsfall in diesem Tabellenfeld existiert, gilt:
$z^u_{p,q,m} = z^o_{p,q,m} = t^o_{p,q,m}$,
sodass der Unzulässigkeitsbereich leer ist.

Gesucht ist eine anonyme Basisdatei, deren Auswertungstabellen den originalen möglichst ähnlich sind, d. h. der Abstand zwischen T^o und T^a sollte minimal sein. Die Ausgestaltung der Funktion zur Messung des Abstandes zwischen T^o und T^a hängt dabei von der konkreten Ausgestaltung des Begriffes der „Tabellenqualität" ab, die sich an dem Bedarf der Datennutzerinnen und -nutzer orientieren sollte.

Für den im Folgenden näher untersuchten Fall der Anonymisierung von ausschließlich kategorialen Merkmalen und damit auch ausschließlich Häufigkeitstabellen als zu kontrollierende Tabellen lässt sich das obige allgemeine SAFE-Problem vereinfachen. Da für jeden einzelnen Datensatz der Mikrodaten in Häufigkeitstabellen nur die Möglichkeit existiert, dass er in einem Tabellenfeld mitgezählt wird oder nicht, lässt sich eine Zuordnungsmatrix A (nur aus 0 und 1 Elementen) bilden. Der Zusammenhang zwischen der Häufigkeit der Sätze in einer Mikrodatendatei und dem Ergebnis der Auswertung in kontrollierten Tabellierungen ist dann:

$$T = AX$$

mit:

X – ist der Häufigkeitsvektor, der angibt, wie oft Objekte mit diesen Merkmalsausprägungen im Datenbestand vorhanden sind. Üblicherweise gilt bei originalen Mikrodaten $x_j = 1$. Werden identische Datensätze bereits vorher zusammengefasst, gilt $x_j > 1$ ($\sum x_j$ = Anzahl der Objekte).

T – Vektor aller Ergebnisse in Auswertungstabellen, also der Vektor aller Tabellenfelder von zu kontrollierenden Häufigkeitstabellen t_i; i = 1, 2, …, k (k-Anzahl der Häufigkeitsfelder in allen zu kontrollierenden Tabellen). Der Vektor hat eine Blockstruktur, wobei jeder Block die möglichen Tabellenfelder genau einer Häufigkeitstabelle der τ Häufigkeitstabellen enthält.

$$T = \begin{Bmatrix} T_1 \\ T_2 \\ \hline T_\tau \end{Bmatrix}$$

A – Zuordnungsmatrix mit $a_{ij} = 1$, wenn das Objekt j im Tabellenfeld i tabelliert wird, sonst $a_{ij} = 0$.

$$A = \begin{Bmatrix} A_1 \\ A_2 \\ \hline A_\tau \end{Bmatrix}$$

mit

$$A_1 = \left\{\begin{array}{ccc} a_{1\,1\,1} = 1 & a_{1\,1j} = 0 & a_{1\,1n} = 0 \\ a_{1\,2\,1} = 0 & a_{1\,2j} = 0 & a_{1\,2n} = 0 \\ \hline a_{1\,i\,1} = 0 & a_{1\,ij} = 1 & a_{1\,in} = 0 \\ \hline a_{1\,(m-1)\,1} = 0 & a_{1\,(m-1)\,j} = 0 & a_{1\,(m-1)\,n} = 0 \\ a_{1\,m\,1} = 0 & a_{1\,mj} = 0 & a_{1\,mn} = 1 \end{array}\right\}$$

Wegen der Blockstruktur im Vektor der Tabellenfelder, die durch die Aneinanderreihung der einzelnen Auswertungstabellen entsteht, gilt auch eine Blockstruktur für A. Die Höhe der Blöcke innerhalb A ist identisch mit der in T und in jedem Block bilden die Spalten Einheitsvektoren, da jedes Objekt nur in genau einem Tabellenfeld einer Auswertungstabelle gezählt wird.

Die in der obigen allgemeinen Schreibweise (1) formulierten Ausschlussintervalle von nicht zulässigen Tabellierungswerten und die Existenz von mindestens drei identischen Einheiten können bei ausschließlich Häufigkeitstabellen einfach durch die folgende Bedingung abgebildet werden:

$$x^a_j = 0, 3, 4, \ldots; \text{ für alle j.}$$

Damit erfüllt dieser Häufigkeitsvektor x^a die Kriterien der k-Anonymität
(d. h. x^a_j ganzzahlig und $x^a_j \neq 1, 2$).

4. Mathematisches Modell und Optimierungsaufgabe

Der verfolgte Ansatz geht von der Bestimmung einer „optimalen Teilmenge" aus dem bereitgestellten Mikrodatenbestand aus. Zur Erhöhung der Schutzwirkung des Verfahrens kann dieser Mikrodatenbestand auch um fiktive Sätze erweitert werden.[4] Für diese Sätze wird die Häufigkeit 0 im originalen Häufigkeitsvektor hinterlegt.

Nach Festlegung der zu kontrollierenden Tabellen lassen sich für den originalen Datenbestand die Vektoren X und T sowie die Zuordnungsmatrix A bestimmen.

Beschreiben:

X^o – Vektor der originalen Häufigkeiten x_i^o der statistischen Objekte der Ausprägungskombination i; i = 1, 2, …, n im Originalbestand (n-Anzahl der Zeilen der Mikrodaten).

T – Vektor aller originalen Tabellenfelder (Häufigkeiten der Objekte) über alle zu kontrollierenden Tabellen t_j; j = 1, 2, …, k (k-Anzahl der Häufigkeitsfelder über alle zu kontrollierenden Randsummentabellen).

A – Zuordnungsmatrix – Blockmatrix mit Einheitsvektoren in den einzelnen zu kontrollierenden Blöcken A_t; t = 1, 2, …, τ (τ-Anzahl der zu kontrollierenden Häufigkeitstabellen). Wenn die Zeile der Mikrodaten j die Ausprägungen so besitzt, das diese Zeile im Tabellenfeld i gezählt wird, gilt $a_{tij} = 1$ sonst $a_{tij} = 0$.

Dann lässt sich der Zusammenhang zwischen den Mikrodaten zu den originalen Tabellenfeldern darstellen als:

$$T^o = AX^o.$$

Bei einer anonymen Datei muss gelten $x_i^a \in \{0, 3, 4, 5, \ldots\}$. Alle vorhandenen Objekte haben eine Häufigkeit von mindestens 3. Die Häufigkeit 0 bewirkt, dass diese Objekte in der anonymen Lösung nicht mehr vorhanden sind.

4 Die Notwendigkeit sollte auf der Grundlage der Gesamtanzahl an Einheiten und der geplanten Auswertungstiefe entschieden werden. Mit diesem Vorgehen wird verhindert, dass man darauf zurückschließen kann, dass jeder Satz der Mikrodaten auch real existieren müsste. Wäre das problematisch, werden die Daten um plausible, aber nicht vorkommende Merkmalskombinationen erweitert.

Für sehr tief gegliederte Auswertungen lässt sich leicht ein Datenbeispiel finden, für das keine Lösung als Gleichungssystem existiert.

Beispiel:

In einer feinen regionalen Gliederung existieren drei Einheiten (z. B. drei Betriebe in einer kleinen Gemeinde).

WZ	Betriebe original	Betriebe anonym	Abweichung
D.10	3	3	±0
D.10.1	1	0	–1
D.10.2	1	3	2
D.10.3	1	0	–1

Diese haben einerseits verschiedene Ausprägungen in den Wirtschaftszweiggruppen, müssen aber auch zusammen als eine Wirtschaftsabteilung (D.10) in Tabellen präsentiert werden können.

Unabhängig davon, wie man die anonyme Lösung gestalten würde, wäre eine absolute Abweichung <2 für alle Tabellenfelder nicht erreichbar. Es lassen sich auch weitere analoge Beispiele zeigen, bei denen eine Abweichung <2 für die Lösung bei mindestens zwei hierarchischen Auswertungstabellen nicht möglich ist. Das obige Gleichungssystem hat somit unter bestimmten Konstellationen keine Lösung $AX^a = T^o$ unter der geforderten Nebenbedingung $x_i^a \in \{0, 3, 4, 5, \ldots\}$. Deshalb ist ein Fehlervektor F (f_j – Fehler im Tabellenfeld j; j = 1, 2, …, k) einzufügen, der die Abweichungen zwischen der Tabellierung der Originaldaten und der anonymen Lösung beschreibt.

Die Menge aller möglichen anonymen Lösungen beschreibt sich dann als

$$\begin{aligned} AX^a + F &= T^o \\ \sum_{i=1}^{n} x_i^a &= O \\ x_i^a &\in \{0,3,4,5,\ldots\} \\ i &= 1,2,\ldots,n \\ j &= 1,2,\ldots,k \end{aligned}$$

mit:

F – Vektor der Tabellierungsfehler, f_j ist die Abweichung des anonymen Ergebnisses zum Original bei der Tabellierung des Tabellenfeldes j,

O – als Gesamtanzahl der statistischen Objekte.

Für die Bestimmung einer eindeutigen Lösung ist zusätzlich eine Zielfunktion Z einzuführen. Die Definition der Funktion orientiert sich an mehreren Zielen:

1. Die Funktion sollte möglichst transparent für die späteren Datennutzerinnen und -nutzer sein. Das Funktionsergebnis sollte für die Interpretation der Datenqualität gut anwendbar sein.
2. Die Funktion sollte innerhalb des Lösungsalgorithmus gut handhabbar sein.
3. Die Funktion sollte sowohl für die in den Tabellen nebeneinander auftretenden großen als auch kleinen Häufigkeiten sinnvolle Optimierungsziele vorgeben.

Vor diesem Hintergrund ist der maximale relative Fehler unbrauchbar, da bei Fallzahlproblemen (Unikaten im Datenbestand, die in einzelnen Tabellenfeldern allein dargestellt werden) ein relativer Fehler von –100 % bzw. +200 % unumgänglich ist. Dieser relative Fehler würde bereits beim Ändern eines geheim zu haltenden Tabellenfeldes von 1 (Unikat) zu 0 bzw. 3 entstehen. Dieser dort mindestens notwendige relative Fehler würde aber zu unbrauchbaren Ergebnissen führen, wenn man ihn für alle Tabellenfelder akzeptieren würde. Eine minimierte Summe der absoluten Abweichungen oder die Summe der Quadrate der Abweichungen erwiesen sich ebenfalls als ungünstig, da bei Testrechnungen einzelne sehr starke Ausreißer nicht verhindert werden konnten. Da für die Datennutzerinnen und -nutzer die Bewertung der Qualität der anonymen Daten für genau eine, ihre jetzt aktuell interessierende Datenabfrage relevant ist, ist die Aussage der mittleren Abweichung oder der mittleren quadratischen Abweichung nur schwer vermittelbar, wenn keine sicheren maximalen Schranken zusätzlich existieren. Der Maximalfehler in den Randsummentabellen erwies sich deshalb als brauchbares Kriterium für die Bestimmung der Optimalität. Dieser zulässige Maximalfehler kann dabei in Abhängigkeit von der Größe des Tabellenfeldes nochmals gestaffelt werden. Mögliche Varianten zur Bestimmung des Vektors g werden in Abschnitt 6 dargestellt.

Es ist somit für die verschiedenen möglichen Lösungsvektoren X die Lösung mit dem kleinsten Maximalfehler gesucht.

$$\begin{aligned} Z &= \min_{X} \left(\max_{j} (|\, f_j \,| - g_j) \right) \qquad (2) \\ AX + F &= T \\ \sum_{i=1}^{n} x_i &= O \\ x_i &\in \{0,3,4,5,\ldots\} \\ i &= 1,2,\ldots,n \\ j &= 1,2,\ldots,k \end{aligned}$$

mit:

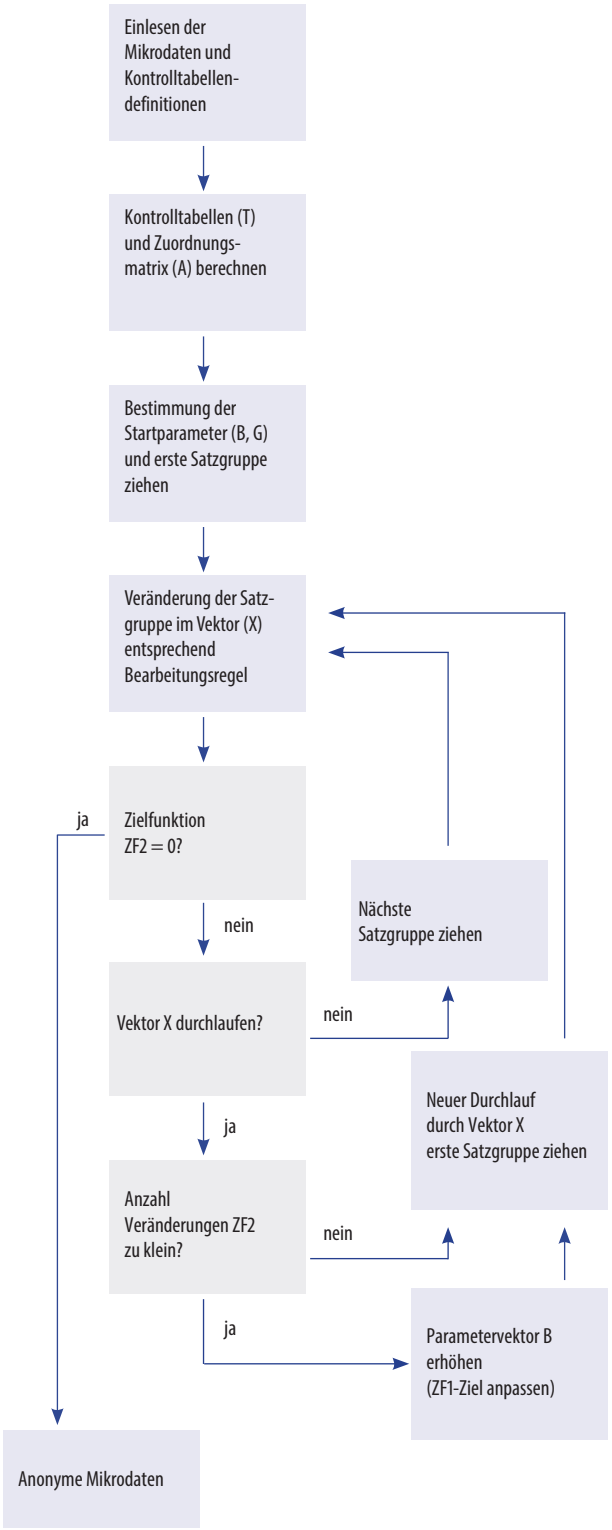
g_j – zusätzlich zulässige Abweichung im Tabellenfeld j, diese zusätzliche Abweichung wird in Abhängigkeit von der Größe des Tabellenfeldes bestimmt, $g_j = G(t_j)$.

Die Aufgabe besteht darin, eine Lösung zu finden, in der keine Merkmalskombinationen mehr auftreten, die bei Auswertungen Geheimhaltungsfälle erzeugen könnten. Deshalb müssen die unerwünschten kleinen Fallzahlen von 1 oder 2 geändert werden (z. B. auf eine Häufigkeit von 3 oder größer bzw. auf die Häufigkeit 0). Die Bedingungen „Erhalt der Gesamtanzahl der Objekte“ und „Minimierung der maximalen Abweichung der Tabellenwerte zwischen Original und Anonymisiert“ sind dabei zu berücksichtigen. Dieses Modell ist somit eine Optimierungsaufgabe mit n+k Unbekannten (Häufigkeiten und Tabellierungsfehler) und k+n+1 Nebenbedingungen (Anzahl Tabellengleichungen, die Mengeneinschränkungen für x_i und die Gesamtanzahl der Objekte). Aufgrund der n Nebenbedingungen zu x_i ist die Aufgabe nichtlinear und ganzzahlig, es können auch mehrere Lösungen unter diesen Bedingungen existieren.

Für reale Datenbestände hat diese Aufgabe eine Dimension, die mit heutiger Rechentechnik nicht explizit lösbar ist. Selbst für ein relativ kleines Beispiel

– 100 000 statistische Objekte mit sieben Merkmalen mit zusammen 16 verschiedenen Schlüsselstufen und 149 817 zu kontrollierenden Tabellenfeldern – ergibt sich ein Speicherbedarf von $(n+k+1)^2$ 4 Byte = 233 GB. Dieser müsste für einen schnellen Zugriff als Hauptspeicher verfügbar sein. Deshalb wurde als Alternative ein performanter numerischer Algorithmus gewählt.

b | Programmablauf „Finden der zulässigen Lösung“



5. Lösungsweg/Numerischer Algorithmus

Numerische Algorithmen zeichnen sich durch eine schrittweise Annäherung an das Optimierungsziel aus. Es wird dabei ein Verfahren festgelegt, das von einer bekannten schlechten Lösung zu einer besseren führt. Die bessere Lösung ist dadurch gekennzeichnet, dass sie näher an der gesuchten Lösung liegt. Das wiederholte Anwenden der Verfahrensregeln führt dann zum Auffinden der gesuchten Lösung.

a) Bestimmung der Startlösung

Da zu Beginn keine anonyme Startlösung bekannt ist, muss diese als erstes bestimmt werden. Dazu gibt es zwei Möglichkeiten:

1. Entweder man bestimmt eine anonyme Startlösung durch einen einfachen Algorithmus. Hier könnte man beispielsweise jeden dritten Satz der Mikrodatendatei mit der Häufigkeit 3 verwenden oder auch andere Algorithmen wählen.
2. Alternativ beschreibt man die Bestimmung der Startlösung als eine separate Optimierungsaufgabe. Dazu „erweitert man die Lösungsmenge“, d. h. es werden auch nicht vollständig anonymisierte Lösungen (für das Ziel eigentlich nicht zulässige Lösungen) in die Lösungsmenge mit aufgenommen. Damit sind die Originalhäufigkeiten des Datenbestandes bereits eine Startlösung. Der Algorithmus muss im ersten Schritt jedoch die Minimierung der Anzahl der noch vorhandenen Geheimhaltungsfälle als primäres Ziel mit in die Zielkriterien aufnehmen, damit der Iterationsalgorithmus das Auffinden der zulässigen Lösung ermöglicht. Der numerische Algorithmus teilt sich dann in die zwei Schritte „Finden der zulässigen Lösung“ und „Optimierung der Lösung“.

Von dieser gefundenen Lösung ausgehend muss dann bei den weiteren Schritten iterativ eine Verbesserung der Qualität der Tabellenfelder erfolgen, wobei jede Veränderung dann so erfolgt, dass keine Geheimhaltungsfälle ($x_j = 1$ oder $x_j = 2$) mehr zugelassen werden.

Nach verschiedenen Tests wurde der zweite Weg verwendet, weil so der Vorteil besteht, dass der Tabellierungsfehler in der gefundenen Startlösung bereits sehr klein ist. Bei Variante 1 sind die entstehenden Tabellierungsfehler zufällig normalverteilt, d. h. es existieren einzelne starke Ausreißer (sehr große Tabellierungsfehler). Bei Variante 2 kann bereits das angestrebte Ziel einer minimalen Maximalabweichung der Tabellierungsfehler als untergeordnetes Zielkriterium mit aufgenommen werden. Die Lösung der Aufgabe wurde damit in zwei Optimierungsaufgaben „Finden der zulässigen Lösung“ und „Optimierung der Lösung“ geteilt, die in zwei Programmen umgesetzt sind.

b) Zielfunktion/Entscheidungsregel/Algorithmus

Der Ablaufplan zur Umsetzung des Algorithmus „Finden der zulässigen Lösung“ ist in Abbildung b dargestellt. Die Zielkriterien für das Verfahren zum „Finden der zulässigen Lösung“ sind die Nachfolgenden, wobei die Nummerierung der Anwendungsreihenfolge entspricht:

1. Minimierung des maximalen Fehlers in den Tabellierungen,
2. Minimierung der Anzahl der verbleibenden Geheimhaltungsfälle in der Datei,
3. Maximierung der Möglichkeiten für weitere Veränderungen,
4. Minimierung des mittleren Fehlers in allen kontrollierten Tabellierungen.

Zwischen dem 1. und dem 2. Ziel besteht ein Widerspruch. Da Originaldaten natürlich einen Tabellierungsfehler von 0 in allen Auswertungen haben, wird ein Verändern der Daten, um Geheimhaltungsfälle zu verringern, zu neuen Tabellierungsfehlern führen. Auch danach ist es nicht immer auszuschließen, dass neue Tabellierungsfehler auftreten. Um diesen Widerspruch aufzulösen, wird das erste Zielkriterium durch die Vorgabe einer fixen Fehlerschranke verfolgt. Diese eröffnet einen Entscheidungsspielraum, der eine Veränderung der Daten zur Verbesserung des zweiten Zielkriteriums erlaubt. Gleichzeitig wird diese Fehlerschranke nur bei unbedingtem Bedarf angepasst. Beginnend mit einem festen Startfehler (in der Regel ±2) wird die Realisierung der Teilziele 2 bis 4 angestrebt. Die Fehlerschranke (bound) von ±2 ist die kleinste mögliche Fehlerschranke (siehe auch Beispiel im Abschnitt 4). Wird mit diesem Startwert für die Fehlerschranke keine Lösung gefunden, so wird sie um 1 erhöht und die Suche fortgesetzt. Wenn Erfahrungen durch die Lösung gleichartiger Beispiele vorliegen, können auch größere Schranken als Startwert vorgegeben werden.

Beim schrittweisen Durchlaufen der Datei werden alle diejenigen potenziellen Veränderungen durchgeführt, die die Ziele 2 bis 4 verbessern. Das bedeutet, es wird folgende Optimierungsaufgabe „Finden der zulässigen Lösung" behandelt:

$$ZF1: \quad \min(\max(b)) \tag{3}$$

$$ZF2: \quad \min\left(\sum_{i=1}^{n} c_i\right)$$

$$ZF3: \quad \min\left(\sum_{i=1}^{k} S\left(\left|t_i^z - t_i^o\right|\right)\right)$$

$$ZF4: \quad \min\left(\sum_{i=1}^{k} \left|t_i^z - t_i^o\right|\right)$$

$$\left|T^z - T^o\right| \le G + B$$

$$T^z = AX^z$$

mit:

$$x_j^z = 0,1,2,3,4,\ldots$$

$$c_j = 1 \quad ; \text{gdw.} \quad x_j \in \{1,2\}$$

$$c_j = 0 \quad ; \text{gdw.} \quad x_j \notin \{1,2\}$$

mit:

X^z – Vektor der Häufigkeiten möglicher Sätze in den Zwischenschritten im Datenbestand
- als Startlösung $X^z = X^o$, wobei
- einmalige Originalsätze ($x_j^o = 1$)
- völlig identische Originalsätze zusammengefasst ($x_j^o > 1$)
- ggf. künstliche Sätze ($x_j^o = 0$)

T^o – Vektor der Häufigkeiten aller Tabellenfelder in den zu kontrollierenden Tabellen bei Auswertung des Originaldatenbestandes, ermittelt als $T^o = AX^o$

A – Zuordnungsmatrix ($a_{ij} = 1$, wenn Objekt j im Tabellenfeld i gezählt wird, sonst $a_{ij} = 0$)

$|T^z - T^o|$ – Vektor aller Tabellierungsfehler

B – Vektor der vorgegebenen (maximal zulässigen/„akzeptierten") Fehlerschranke (bound) für die einzelnen Tabellenfelder. b_i ist der zulässige Fehler im Feld i[5].

G – Vektor eventuell erlaubter zusätzlicher Fehler aufgrund der Größe der Tabellenfelder[6]

C – Vektor der Geheimhaltungsfälle $c_i = 1$, wenn x_i eine Häufigkeit von 1 oder 2 hat, sonst $c_i = 0$.

S() – Straffunktion für zu kleine Randabweichungen. Entspricht die Abweichung eines Tabellenfeldes der erlaubten Maximalabweichung, so kann kein Satz mehr in diese Richtung geändert werden, ohne dass eine sofortige Kompensation innerhalb der Satzgruppe erfolgen muss (bei $t_i = 3$ und $b = 3$ ist ein weiteres Erhöhen blockiert, da dann $t_i = 4$ die bounds b verletzt). Es wird versucht, diese Konstellation zu vermeiden. Deshalb sind Tabellenfelder mit Abweichungen, kleiner als der zulässige Maximalfehler, günstiger und werden durch diese Straffunktion bevorzugt. Für die genaue Ausgestaltung dieser Funktion siehe Abschnitt 6a.

Veränderungen werden nur bei Einhaltung der gegebenen bounds b (ZF1) so vorgenommen, dass sie die Anzahl an Geheimhaltungsfällen minimieren (ZF2). Die Lösung der Aufgabe ist erreicht, wenn die Summe der $c_i = 0$ ist, also kein x_i mehr den Wert 1 oder 2 besitzt. Sollten bezüglich der Anzahl der Geheimhaltungsfälle neutrale Veränderungsmöglichkeiten existieren, so wird das Qualitätsziel 3 (ZF3) und bei Neutralität von 2 und 3 das Qualitätsziel 4 (ZF4) für die Entscheidung berücksichtigt.

Vor der Bearbeitung der Mikrodatendatei werden die Sätze so sortiert, dass möglichst wenige Schlüsseländerungen zwischen benachbarten Zeilen anzutreffen sind. Durch die Sortierung des Datenbestan-

5 Die Werte b_i können im Programm getrennt für ein- und mehrdimensionale Tabellenfelder verwaltet werden, wobei für mehrdimensionale Tabellenfelder nur ein fester zusätzlicher Abweichungswert für alle Felder gilt. Damit sind die b_i zwei einheitliche Werte Parameter für alle i (zwei Programmparameter).

6 Die Gewichtung der Tabellenfelder erfolgt nach der Größe des Tabellenwertes $g_i = G(t_i)$. Großen Tabellenfeldern wird so eine größere zulässige Abweichung erlaubt. Ein Beispiel implementierter Gewichtung ist $g_i = \text{int}(\log_{10} t_i)$. Damit werden folgende zusätzliche Abweichungen zulässig:

Tabellenwerte	1 bis 9	10 bis 99	100 bis 999	1 000 bis 9 999	…
zusätzliche Abweichung	0	1	2	3	

Es sind aber auch andere Gewichtungsfunktionen möglich (siehe Abschnitt 6f).

des befinden sich Datensätze nebeneinander, die in vielen Auswertungen im gleichen Tabellenfeld tabelliert werden. Bei der gleichzeitigen Behandlung von drei Unikaten würde durch die Veränderung von zwei Datensätzen zu 0 und einem zu 3 in diesen Tabellenfeldern eine Kompensation der Veränderungen auftreten. Um diese Kompensationseffekte bei der Veränderung benachbarter Sätze mit ausnutzen zu können, werden sequenziell gleitende Gruppen von drei, ggf. vier Sätzen aus der sortierten Datei gezogen. Für diese Gruppen werden alle möglichen Veränderungskombinationen ±1, ±2 und ggf. ±3 gebildet. Die Auswahl einer durchzuführenden Veränderung erfolgt dann nach folgendem Schema:

1. Ermittlung aller Veränderungskombinationen ±1, ±2 und ggf. ±3, für die die Häufigkeit der einzelnen Merkmalskombinationen nichtnegativ bleibt ($x_j >= 0$). Die Veränderung +2 kommt bei $x_j = 1$ und die Veränderung –2 bei $x_j = 2$ zur Anwendung, damit eine nicht anonyme Häufigkeit in beide Richtungen direkt auf anonym (d. h. 0 oder 3) geändert werden kann. Die Veränderung +3 kommt bei $x_j = 0$ und die Veränderung –3 bei $x_j = 3$ zur Anwendung, damit keine anonyme Häufigkeit auf 1 oder 2 (d. h. nicht anonym) zurückgeändert werden muss.
2. Für diese Veränderungskombinationen wird die Verletzung des vorgegebenen Maximalfehlers getestet und die Veränderung der Ziele 2–4 ausgewertet. Es werden dann nur noch die Veränderungskombinationen ausgewählt, die die Maximalfehlerbeschränkung nicht verletzen (keine Verletzung der bounds in ZF1 zulassen).
3. Die Menge aller Veränderungskombinationen wird reduziert um die Kombinationen, die nach ihrer Realisierung mehr Geheimhaltungsfälle erzeugen würden, als vorher vorhanden waren (keine Verschlechterung von ZF2 zulassen).
4. Die durchzuführende Veränderung wird bestimmt, indem man die Kombinationen auswählt, die das Ziel 2 (die Geheimhaltungsfälle zu beseitigen) am meisten verbessert. Verbleiben mehrere Kombinationen in der Auswahl, so wird das nächste Teilziel (erst Ziel 3, dann Ziel 4) für die Entscheidung herangezogen. Es werden ggf. auch die bezüglich höherwertiger Teilziele neutralen Kombinationen ausgewählt, wenn keine Kombination das höherwertige Teilziel verbessert. Eine ausgewählte Kombination verbessert somit mindestens ein Teilziel. Verschlechterungen eines Teilzieles sind nur bei gleichzeitiger Verbesserung eines höherwertigen Zieles möglich.

Ist die Ergebnismenge leer, findet keine Veränderung statt. Ansonsten wird die erste Kombination aus der Menge durchgeführt.

Beispiel:

Es ist eine Datei mit den qualitativen Merkmalen Wirtschaftszweigklassifikation (WZ) und Region zu anonymisieren. Als zu kontrollierende Randsummentabellen seien nur folgende eindimensionale Randsummen zu testen: die eindimensionalen Tabellierungen der Wirtschaftszweigklassifikation als 2-, 3- und 5-Steller und die Region für Berlin nach Stadtteil (Berlin-West, Berlin-Ost) und Bezirk. Mehrdimensionale Tabellen seien vernachlässigt. Die unteren Tabellen enthalten die Abweichungen in der Anzahl der Betriebe, die sich im Verlauf der bisherigen Anonymisierung ergeben haben.

Tabellenfelder, in denen keiner der Sätze der ausgewählten Gruppe tabelliert wird, können durch die Kombinationen auch nicht verändert werden. Deshalb sind nur die grau hinterlegten Tabellenfelder der folgenden Tabellen zu betrachten.

WZ 2-Steller	Abw.
…	…
CA	–1
CB	2
DA	2
DB	–1
DC	±0
…	…

WZ 3-Steller	Abw.
…	…
DA1	3
DA2	–1
DB1	±0
DB2	–1
DC1	±0
…	…

WZ 5-Steller	Abw.
…	…
DA189	1
DA191	1
DA196	–1
DA197	1
DA200	±0
…	…

Stadtteil	Abw.
Berlin-Ost	1
Berlin-West	–1

Bezirke	Abw.
…	…
03	1
04	–2
05	2
06	–1
…	…

Die aktuell zulässige Maximalfehlerschranke (bound b) sei ±3. Eine weitere Unterscheidung der Tabellenabweichung g_i sei vernachlässigt ($g_i = 0$ für alle Tabellenfelder). Folgende Satzgruppe des Datenbestandes wird aktuell untersucht:

Satz in der Satzgruppe	WZ	Bezirk	Stadtteil Berlin-…	Anzahl Betriebe	Veränderungsvarianten der Zeile
…	…	…	…	…	
1	DA191	03	Ost	3	–1, +1, –3
2	DA197	05	West	1	–1, +1, +2
3	DA200	06	West	1	–1, +1, +2
…	…	…	…	…	

Damit ergeben sich folgende Veränderungsvarianten:

Zielkriterien/Entscheidungskriterien:
1*) Bleibt Veränderung innerhalb der maximalen Fehlerschranke?
2*) Veränderung der Anzahl an Geheimhaltungsfällen,
3*) Veränderung des Randabstandes der Lösung (Bei diesem Beispiel sei vereinfacht die Anzahl der Tabellenfelder mit $|t_i| = b$ als ZF3 betrachtet),
4*) Veränderung der Summe der absoluten Randsummenfehler.

Nach Entfernung der Veränderungskombinationen, bei deren Realisierung die zulässigen Tabellenabweichungen überschritten werden (ZF1), werden auch die Kombinationen entfernt, deren Umsetzung mehr Geheimhaltungsfälle als vorher erzeugen würde (Spalte ZF2). Beide sind in Tabelle 1 grau unterlegt. Aus den verbleibenden Möglichkeiten wird nach folgender Auswahlregel gewählt (jeweils fett dargestellt):

1. Die meisten Geheimhaltungsfälle beseitigen die Kombinationen 17, 33 und 35 (jeweils 2).
2. Die größte Verbesserung des Randabstandes (oder geringste Verschlechterung) erzeugt darunter die Kombination 35. Wären hier immer noch mehrere Kombinationen gleichwertig, entscheidet der Einfluss auf den mittleren Randsummenfehler (Spalte ZF4 möglichst klein).

Die vorgeschlagenen Veränderungen der Kombination 35 werden durchgeführt. Bei diesem Beispiel ist erkennbar, dass ggf. auch nicht geheim zu haltende Fälle mit verändert werden, wenn es für das Gesamtproblem nützlich ist. Diese Veränderung von nicht geheim zu haltenden Fällen ermöglicht beispielsweise die Revision von bereits getroffenen Änderungsentscheidungen, aber auch die Bildung größerer anonymer Gruppen als 3. Sie werden aber nur eingeschränkt geprüft, z. B. wenn sonst keine Lösung möglich ist (siehe weiter unten „Auswahltechniken der zu testenden Satzgruppen").

Danach wird eine neue Satzgruppe von x Sätzen aus der Datei gezogen, die nach den gleichen Entscheidungsregeln getestet und bearbeitet wird. Dieser Algorithmus „Auswahl einer Gruppe von Sätzen und Entscheidung der Veränderung" wird ständig wiederholt. Da eine Veränderung nur bei Annäherung an das Ziel (siehe Zielfunktionen) durchgeführt wird, können keine Schleifen auftreten. Die Veränderung zu einer alten Zwischenlösung wäre nur bei einer Verletzung der obigen Entscheidungsregeln möglich.

Aus kombinatorischer Sicht wäre es auch möglich, den ersten Satz unverändert zu lassen und nur den zweiten und/oder dritten Satz zu ändern. Diese Kombinationen werden aber nicht geprüft, da sie in der folgenden Satzgruppe mit enthalten sind und dort aber zusätzlich die Kompensation mit den nachfolgenden Sätzen geprüft wird.

1 | Beispiel für Entscheidungsregeln

Variante	Veränderung Satz ... um …			Veränderung Zielkriterien				Bemerkung
	1	2	3	ZF1*)	ZF2*)	ZF3*)	ZF4*)	
1	–1	–1	-1	nein	–1	0	–3	
2	–1	–1	1	nein	0	–1	–10	
3	–1	–1	2	nein	–1	–1	–6	
4	–1	–1	0	nein	0	–1	–8	
5	–1	1	–1	nein	0	1	1	
6	–1	1	1	nein	1			
7	–1	1	2	ja				
8	–1	1	0	nein	1			
9	–1	2	–1	ja				
10	–1	2	1	ja				
11	–1	2	2	ja				
12	–1	2	0	ja				
13	–1	0	–1	nein	0	–1	–2	
14	–1	0	1	nein	1			
15	–1	0	2	nein	0	0	–1	
16	–1	0	0	nein	1			
17	1	–1	–1	nein	**–2**	**1**	5	
18	1	–1	1	nein	–1	1	1	
19	1	–1	2	ja				
20	1	–1	0	nein	–1	0	2	
21	1	1	–1	ja				
22	1	1	1	ja				
23	1	1	2	ja				
24	1	1	0	ja				
25	1	2	–1	ja				
26	1	2	1	ja				
27	1	2	2	ja				
28	1	2	0	ja				
29	1	0	–1	ja				
30	1	0	1	ja				
31	1	0	2	ja				
32	1	0	0	ja				
33	–3	–1	–1	nein	**–2**	**1**	4	
34	–3	–1	1	nein	–1	–1	–4	
35	–3	–1	2	nein	**–2**	**–1**	**–3**	zu wählende Kombination
36	–3	–1	0	nein	–1	–1	–1	
37	–3	1	–1	nein	–1	0	4	
38	–3	1	1	nein	0	0	0	
39	–3	1	2	nein	–1	0	5	
40	–3	1	0	nein	0	0	–1	
41	–3	2	–1	ja				
42	–3	2	1	ja				
43	–3	2	2	ja				
44	–3	2	0	ja				
45	–3	0	–1	nein	–1	–1	3	
46	–3	0	1	nein	0	–1	–5	
47	–3	0	2	nein	–1	–1	0	
48	–3	0	0	nein	0	–1	–2	

6. Aspekte des Algorithmus

a) Maximierung der Möglichkeiten für weitere Veränderungen (Teilziel ZF3)

Bei einer Auswahl aus mehreren Veränderungsmöglichkeiten, die die gleiche Anzahl an Geheimhaltungsfällen beseitigen, wird zuerst der Einfluss auf die Möglichkeiten für weitere Veränderungen berücksichtigt. Innerhalb der Satzgruppe können sich in der Regel nicht alle Veränderungen in den Tabellierungsfehlern gegenseitig kompensieren, da die Sätze nicht vollständig identisch sind. Eine einzelne Datensatzveränderung ist nur dann gefährdet, wenn die Fehler in den Tabellenfeldern sich auf beiden Seiten (positive und negative Abweichungen) gleichzeitig zu dicht am Rand der zulässigen Abweichung befinden. Für die Beseitigung eines Geheimhaltungsfalles durch Verringerung der Häufigkeit um 1 darf kein Tabellenfeld i einen negativen Fehler haben, der gleich dem zulässigen Maximalfehler $-(b_i+g_i)$ ist. Analog ist die Beseitigung eines Geheimhaltungsfalles durch Erhöhung der Häufigkeit von 1 auf 3 nur dann möglich, wenn die Abweichung des Tabellenfeldes nach oben weder (b_i+g_i) noch $(b_i+g_i)-1$ beträgt. Weiterhin verhindert eine Tabellenfeldabweichung von $>=(b_i+g_i)-2$ bzw. $<=2-(b_i+g_i)$ die Korrektur einer Geheimhaltungsentscheidung in Form eines Wechsels der Häufigkeit von 0 auf 3 oder umgekehrt. Deshalb werden diese Abweichungen in Tabellenfeldern im Teilziel ZF3 als „kritische Abweichungen" berücksichtigt, wenn eine Auswahlmöglichkeit unter mehreren Kombinationen besteht, die die gleiche Anzahl an Geheimhaltungsfällen beseitigen.

Eine einfache Fehlerfunktion zum „mittleren" Tabellierungsfehler (wie ZF4) kann dem Ziel „Maximierung der Möglichkeiten für weitere Veränderungen im Datenbestand" nicht gerecht werden, vor allem dann nicht, wenn auch noch versucht wird, mit zwei getrennten Maximalfehlerschranken für ein- und mehrdimensionale Tabellen zu arbeiten. Als Regel wird nebenstehende Straffunktion S zur Messung des Randabstandes verwendet.

$$S=\sum_{i=1}^{k} s_i$$

$$s_i=\begin{cases} 9 & ;\forall\,|f_i|=b_i+g_i \\ 4 & ;\forall\,|f_i|=b_i+g_i-1 \\ 1 & ;\forall\,|f_i|=b_i+g_i-2 \\ 0 & ;\forall\,|f_i|<b_i+g_i-2 \end{cases}$$

Die Minimierung dieser Straffunktion steht als Ziel somit vor der allgemeinen Verbesserung der Summe der Tabellierungsfehler, weil sie sich positiv auf die Wahrscheinlichkeit auswirkt, dass weitere Veränderungen möglich sind.

b) Numerische Probleme

Bei numerischen Algorithmen können mehrere Probleme auftreten:

1. Schleifen

   Um zu verhindern, dass der Algorithmus sich in einer endlosen Schleife aufhängt, bestehen einige Anforderungen an die Auswahlregeln. Die Funktionen, mit deren Hilfe die Auswahl getroffen wird, müssen folgende beiden Bedingungen erfüllen:

- Für zwei beliebige Lösungen X_1 und X_2 des Lösungsraumes gilt: Der Abstand von X_1 nach X_2 ist gleich dem negativen Abstand X_2 nach X_1

$$\overline{X_1X_2}=-\overline{X_2X_1}$$

Diese Bedingung würde beispielsweise dann vernachlässigt, wenn man versuchte, sich bei der Straffunktion nur auf die störenden Tabellenfelder eines gerade in der Auswahlgruppe betrachteten Geheimhaltungsfalls zu konzentrieren. Eine „Bevorzugung" eines aktuellen Geheimhaltungsfalles führt bei einem Wechsel zum nächsten Geheimhaltungsfall automatisch dazu, dass die obige Regel verletzt ist und somit Schleifen nicht mehr ausgeschlossen werden können. Im obigen Fall hat jede Lösung für alle drei Zielkriterien immer genau einen Funktionswert (unabhängig von der Veränderungsrichtung). Damit gilt:

$$\overline{X_1X_2}=f(X_2)-f(X_1)=-(f(X_1)-f(X_2))=-\overline{X_2X_1}$$

- Die Zielkriterien müssen einerseits in Optimierungsrichtung beschränkt sein und andererseits sicherstellen, dass es eine Lösungsmenge gibt, die die gesuchte Lösung enthält. Die zweite Teilfunktion (ZF2 – Anzahl der vorhandenen Geheimhaltungsfälle) ist in Optimierungsrichtung (nach unten) beschränkt, denn es können nicht mehr Geheimhaltungsfälle entfernt werden, als vorhanden sind. Gleichzeitig ist das Minimum (0 Geheimhaltungsfälle) identisch mit der gesuchten Lösung. Die dritte Teilfunktion (ZF3 – Straffunktion für in Nähe des Maximalfehlers liegende Tabellenfelder) ist ebenfalls nach unten beschränkt. Sie ist 0, wenn kein Tabellenfeld einen Wert im Bereich $f_{max} \geq abs(f_i) \geq f_{max}-2$ besitzt. Die vierte Teilfunktion (ZF4 – Summe der Fehler in Tabellenfeldern) hat ihr Minimum, wenn alle Tabellenfelder fehlerfrei sind. Da für die dritte und vierte Teilfunktion keine Einschränkungen bezüglich Zulässigkeit gestellt werden, sind alle Kombinationen bezüglich dieser Teilfunktionen zulässige Lösungen. Die Zielfunktion ZF1 ist ebenfalls nach unten beschränkt. Der kleinste absolute Fehler kann nur 0 sein. Es besteht aber das Problem, dass nicht sicher ist, ob bei einem Fehler von 0 eine Lösung existiert. Die Bestimmung einer Lösungsmenge, die die Lösung enthält, wird im Rahmen der im folgenden dargestellten Stagnation gelöst.

2. Stagnation

   Es kann vorkommen, dass zwar einerseits der Algorithmus zum Ziel konvergiert, andererseits aber die Geschwindigkeit so langsam ist, dass man nicht in einer vertretbaren Zeit zum Ergebnis kommt. In diesem Fall muss der Algorithmus die Situation erkennen können und reagieren. Das wird bei diesem Verfahren mit dem „Gashebel" zulässiger Maximalfehler (bound) geregelt. Wird bei einem Durchlauf nicht ein erwarteter Anteil von Geheimhaltungsfällen beseitigt, so erfolgt eine Vergrößerung der bounds. Damit entstehen wieder größere Freiräume, wodurch beim erneuten Durchlauf weitere Geheimhaltungsfälle beseitigt werden können. Die Geschwindigkeit des Verfahrens lässt sich somit über den Parameter „Anteil der mindestens zu beseitigenden Geheimhaltungsfälle" regeln. Eine zu große Geschwindigkeit geht hierbei jedoch zu Lasten der Qualität des Ergebnisses (höherer Maximalfehler in der Lösung).

Dass es bounds geben muss, für die eine Lösung existiert, kann man mit folgendem Beispiel zeigen. Verwendet man eine vereinfacht generierte Startlösung wie beispielsweise jeden dritten Satz des Datenbestandes mit der Häufigkeit 3, so lassen sich auch aus diesem Bestand alle Kontrolltabellen berechnen. Verwendet man die aus diesen Kontrolltabellen bestimmbare Maximalabweichung als bound, so ergibt sich ein Lösungsraum an möglichen Datenbeständen, für den mindestens eine zulässige Lösung existiert (die generierte Triviallösung). Auch wenn diese Lösung mit Sicherheit nicht die gesuchte Lösung ist, so genügt sie dem Existenzbeweis einer Lösung, der für den Nachweis der Lösbarkeit erforderlich ist.

c) Auswahltechniken der zu testenden Satzgruppen

Die Entscheidung zur Beseitigung der Geheimhaltungsfälle durch Verändern der Häufigkeit wird nur dann verhindert, wenn die Veränderung der Häufigkeit mit einer Verletzung der Schranke der erlaubten Maximalabweichung (b_i+g_i) in einem Tabellenfeld einhergehen würde. Dann wird die Entscheidung verschoben. Durch nachfolgende Veränderungen kann es durchaus sein, dass bei einem erneuten Testen der gleichen Satzgruppe die dann vorhandenen Abweichungen im Tabellenfeld die Beseitigung ermöglichen, da sich die konkreten Abweichungen mit jeder Veränderung im Datenbestand ebenfalls verändern können und auf die Schaffung von entsprechenden Freiräumen Wert gelegt wurde. Außerdem können ggf. Teile der Satzgruppe gruppiert mit nachfolgenden Sätzen anonymisierbar sein.

Durch gezielte Auswahltechniken sind folgende Probleme zu lösen:

- Gruppierung zu Häufigkeiten größer 3,
- automatische Erkennung der erforderlichen bounds,
- Kontrolle und ggf. Korrektur von „alten" Geheimhaltungsentscheidungen.

Zu Beginn eines Anonymisierungslaufs ist die erforderliche Maximalabweichung in der Regel nicht bekannt. Bei mehrfachen Läufen für den gleichen Datenbestand (z. B. Monatsreihen) könnten ggf. Erfahrungen vorliegen. Es muss aber ein Startwert der bounds vorgegeben werden. Problematisch sind dabei zu eng gestellte bounds. Das würde dazu führen, dass zuerst nur die homogenen Satzgruppen anonymisiert (zusammengefasst) werden, und im Nachhinein immer inhomogenere Sätze übrig bleiben, denn die gleichartigen Satzgruppen dazwischen wurden entfernt. Bei einem ständigen Durchsuchen der Gesamtdatei würden mit großem Testaufwand die homogen Satzgruppen anonymisiert, während die Anonymisierung der inhomogenen Sätze immer weiter verschoben werden. Um die erforderlichen bounds automatisch und schnell durch den Algorithmus zu erkennen, wird folgendes Verfahren verwendet:

Es wird nur ein kleiner Teil der Datei bearbeitet. Für diesen Teil wird die Anonymisierung durchgeführt. Erst, wenn dieser Teil fast vollständig abgearbeitet ist, wird ein weiterer Teil der Gesamtdatei dazu genommen. Das „fast vollständig" ist erforderlich, um strukturelle Ausreißer mit den größeren kombinatorischen Möglichkeiten der Gesamtdatei entscheiden zu können. Ist die Anonymisierung für diesen Teil nicht möglich, wird bereits der bound-Parameter im Modell höher gesetzt. Eine Erhöhung des Parameters schafft in jedem Fall neue Freiräume, um weitere Veränderungen an geheim zu haltenden Sätzen vorzunehmen. Somit kann auch bei sehr großen Dateien schnell der notwendige bound-Parameter erkannt werden und das Verfahren bedeutend schneller laufen. Die verbleibenden Reste des bereits bearbeiteten Teilbestandes werden regelmäßig wieder mitgetestet, da sich durch jede Veränderung auch die Tabellierungsabweichungen ändern und somit ständig andere Möglichkeiten existieren.

Beim Durchlaufen der Datei werden zuerst standardmäßig nur die noch existierenden Geheimhaltungsfälle betrachtet. Es werden immer drei benachbarte Geheimhaltungsfälle getestet. Tritt trotzdem eine Stagnation des Verfahrens ein, so werden nacheinander andere Auswahltechniken durchgeführt, wobei nach jedem Lauf getestet wird, ob die Stagnation noch vorhanden ist. Die weiteren Auswahltechniken sind:

1. Es werden alle noch vorhandenen Sätze (Häufigkeit > 0) getestet. Da es auch vorkommen kann, dass in der Nähe eines Geheimhaltungsfalles (zwei Sätze davor und dahinter) alle Sätze entfernt wurden, werden im ersten Versuch alle nicht mehr vorhandenen Sätze (Häufigkeit = 0) vernachlässigt und die Auswahlgruppen aus der verbleibenden Menge gebildet. Dieses Verfahren zeichnet sich auch dadurch aus, dass mit kleinen Veränderungen der Häufigkeit bei bereits anonymisierten Sätzen (±1) auch die Teilziele 3 und 4 verbessert werden können. Durch diese Gruppenauswahl wird beispielsweise die Möglichkeit, einen Geheimhaltungsfall zu einer bereits vorhandenen 3er-Gruppe mit hinzuzunehmen, getestet. Es werden dann vier identische Einheiten gebildet.
2. Die noch vorhandenen Geheimhaltungsfälle werden zusammen mit ihren im sortierten Bestand physisch benachbarten Sätzen getestet. Das können auch bereits durch die Anonymisierung entfernte Sätze sein, wobei die Entscheidung für diese benachbarten Sätze revidiert werden kann.
3. Als letzte Variante kann auch ein Gesamtdurchlauf der Datei durchgeführt werden. Dann könnte auch das Entfernen eines Satzes bei allen Sätzen noch einmal revidiert werden, wenn es dem Gesamtziel (siehe Entscheidungsregeln) dient. Dieses Auswahlverfahren ist am rechenaufwändigsten und wird deshalb nur dann angewendet, wenn bereits der Gesamtbestand in den Anonymisierungslauf einbezogen ist und nur noch sehr vereinzelt Geheimhaltungsfälle existieren.

Da die Anzahl der zu ziehenden Satzgruppen und auch die Anzahl der Änderungsmöglichkeiten in den Satzgruppen bei den verschiedenen Auswahltechniken unterschiedlich ist, wird auch die Erwartung bezüglich der Anzahl der zu beseitigenden Geheimhaltungsfälle für die Auswahltechniken verschieden

angesetzt. Auswahltechniken, bei denen sehr viele Satzgruppen gebildet und getestet werden, müssen auch entsprechend mehr Geheimhaltungsfälle beseitigen, um noch als performant zu gelten.

d) Zeitverlauf/Zeitbedarf

Das Verfahren zeichnet sich durch einen hyperbolischen Zeitverlauf aus. Einer sehr schnellen Anonymisierung der ersten Hälfte folgt eine sehr langsame Anonymisierung der zweiten Hälfte der Merkmalsträger. Nach ca. 50 % der Gesamtlaufzeit sind in der Regel nur noch weniger als 5 % der Geheimhaltungsfälle zu anonymisieren. Für diese wird wegen der Nutzung von aufwändigen Auswahlalgorithmen mehr Zeit benötigt. Es wäre durch Anpassung des Stagnationsmaßes auch möglich, den Zeitverlauf linearer zu gestalten. Das geht aber mit einer Erhöhung der Maximalabweichung im Ergebnis einher (bei Tests ergab sich eine ca. 25 % größere Maximalabweichung).

e) Eindimensionale Tabellierungen besser erhalten

Die Wertigkeit eindimensionaler Tabellenfelder wird meist höher eingeschätzt als die mehrdimensionaler Tabellenfelder. So ist Datennutzerinnen und -nutzern beispielsweise die Anzahl der Einwohner insgesamt in einer Region wichtiger als die Häufigkeiten in komplexeren mehrdimensionalen Auswertungen. Es wurde von Datenproduzentinnen und -produzenten sowie Nutzerinnen und Nutzern die externe Forderung postuliert, dass diese Tabellenfelder einen geringeren Fehler im Ergebnis aufweisen müssen. Dabei wurde als Zielvorstellung für eindimensionale Auswertungen der theoretische Minimalfehler von ±2 angestrebt. Für weniger als 2 wurden zuvor bereits Gegenbeispiele gezeigt, bei denen die Forderung nicht realisierbar ist.

Beim SAFE-Programm können deshalb unterschiedliche Fehlerschranken ein- und mehrdimensional vorgegeben werden. Da eine ständige Restriktion von 2 jedoch für das numerische Verfahren zu eng ist, wurde folgende Regel eingebaut. Es können dem Programm drei Parameter übergeben werden. Die Parameter haben folgende Bedeutung:

1. Startfehler eindimensionaler Tabellenfelder,
2. Startfehler mehrdimensionaler Tabellenfelder,
3. maximaler Abstand der beiden Fehlerschranken.

Bei Stagnation und der Entscheidung, die bounds b zu vergrößern, werden so lange nur die b_i mehrdimensionaler Tabellenfelder um 1 erhöht, bis der maximale Abstand erreicht ist, danach erfolgt ein gleichzeitiges Erhöhen aller b_i um 1.

f) Zusätzliche Gewichtung bei großen Zellwerten

In den Tabellen sind bei großen Zellwerten größere Abweichungen in den Tabellenfeldern durch den eingeführten Fehlervektor g zulässig. Eine größere Abweichung bedeutet bei einem großen Tabellenwert eine kleinere relative Veränderung und kann deshalb dort eher akzeptiert werden. In der Anwendung von SAFE werden verschiedene Regeln für Größenklassen verwendet. Für alle gilt: Je größer die Größenklasse, desto größer ist der zulässige zusätzliche Fehler.

Während in der ersten Version (Variante 1) die Gewichte im Fehlervektor g nach dem dekadischen Logarithmus gebildet wurden, ist aktuell auch eine zweite Gewichtungsversion im Einsatz. Bei beiden aktuellen Versionen gibt es zehn Größenklassen. Die dekadischen Größenklassen entsprechend dem dekadischen Logarithmus, werden also als $g_i = \text{int}(\log_{10} t_i)$ gebildet (Tabelle 2, Variante 1). Die größte Größenklasse enthält Tabellenfelder mit einer Besetzung von $>= 1000\,000\,000$. Bei der zweiten Version gibt es im Bereich der Zellbesetzungen zwischen 10 und 1000 Einheiten noch weitere Untergliederungen (Tabelle 2, Variante 2). Beide Gewichtungsversionen wurden bereits mehrfach getestet, die Version mit den zusätzlichen Größenklassen bei Besetzungszahlen unter 1000 wird derzeit bevorzugt, da in den „kleineren" Tabellenfeldern, bei denen eine Abweichung um 1 noch größere Auswirkungen auf den relativen Fehler hat, die Abweichungen noch stärker minimiert werden.

7. Korrektur der Lösung

Um dem Wunsch nach einem möglichst kleinen Maximalfehler stärker nachzukommen, wird im Anschluss an die Anonymisierung noch folgende Korrekturaufgabe gelöst. Ziel ist es, sowohl die eindimensionalen Tabellenfelder stärker zu verbessern als auch die unterschiedliche Gewichtung schrittweise aus der Lösung zu entfernen. Der Ablaufplan zur Softwareumsetzung der Aufgabe ist in Abbildung c dargestellt.

Dazu wird die Optimierungsaufgabe „Optimierung der Lösung" gelöst:

2 | Tabellenwerte der Gewichtungsversionen

zusätzliche Abweichung	Variante 1		Variante 2	
0	1 bis	9	1 bis	9
1	10 bis	99	10 bis	19
2	100 bis	999	20 bis	49
3	1 000 bis	9 999	50 bis	99
4	10 000 bis	99 999	100 bis	199
5	100 000 bis	999 999	200 bis	999
6	1 000 000 bis	9 999 999	1 000 bis	9 999
7	10 000 000 bis	99 999 999	10 000 bis	99 999
8	100 000 000 bis	999 999 999	100 000 bis	999 999
9	1 000 000 000 und mehr		1 000 000 und mehr	

$$
\begin{aligned}
&\text{ZF 1:} \quad \min\left(\max(b)\right) \\
&\text{ZF 2:} \quad \min\left(\sum_{i=1}^{n} c_i\right) \\
&\text{ZF 3:} \quad \min\left(\sum_{i=1}^{k} S_K\left(\left|t_i^z - t_i^o\right|\right)\right) \\
&\text{ZF 4:} \quad \min\left(\sum_{i=1}^{k} \left|t_i^z - t_i^o\right|\right) \qquad \text{mit:} \\
&\left|T^z - T^o\right| \le F \qquad\qquad x_j^z = 0,3,4,\ldots \\
&T^z = AX^z \qquad\qquad c_j = 1 \;\; ;\text{gdw.} \;\; t_j^z > f_j \\
&\qquad\qquad\qquad\qquad\quad c_j = 0 \;\; ;\text{gdw.} \;\; t_j^z \le f_j
\end{aligned}
\tag{4}
$$

mit:

C – Vektor der Schrankenverletzungen
$c_i = 1$, wenn im Tabellenfeld t^z_i die Abweichung den zulässigen Wert f_i übersteigt, sonst $c_i = 0$.

Alle anderen Variablen behalten ihre Bedeutung.

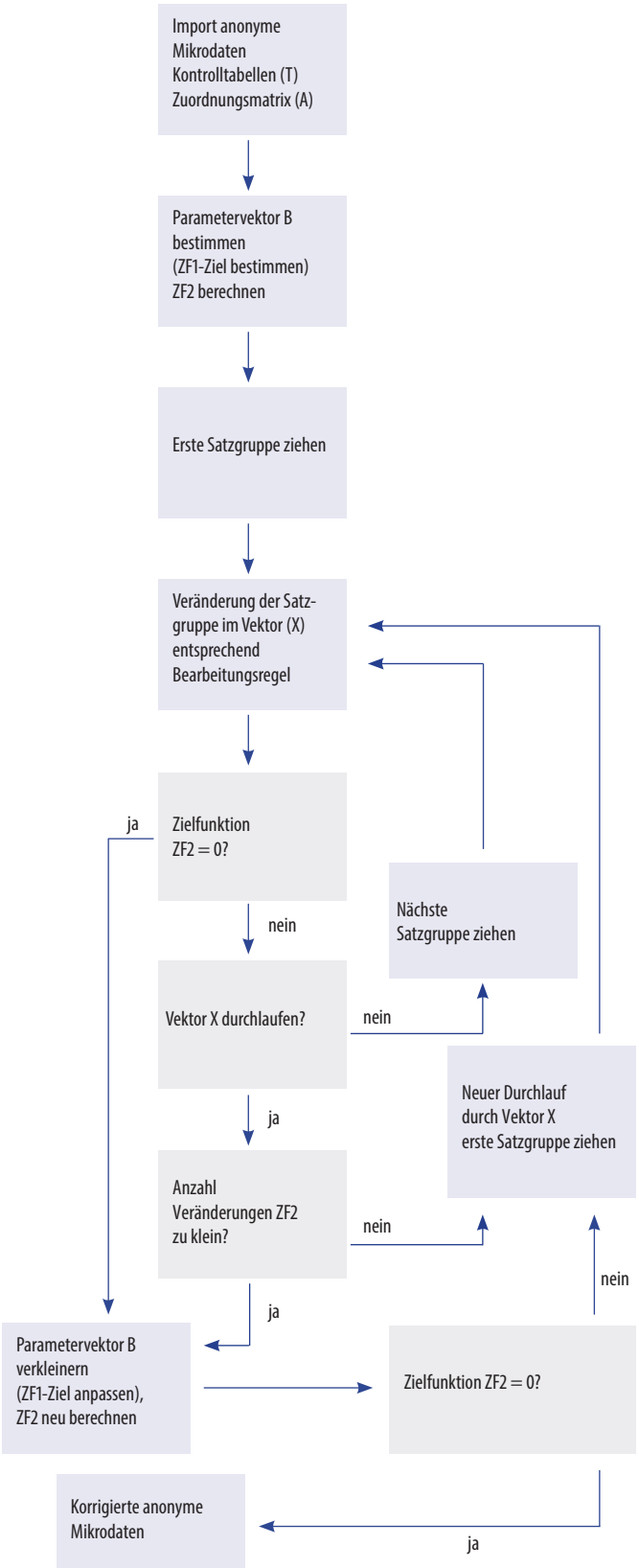
Im Unterschied zur Aufgabe 3, bei der keine Schrankenverletzungen erlaubt waren, wird in dieser Aufgabe d gezielt die Verletzung einzelner Schrankenwerte beseitigt. Diese Schrankenverletzung kann erzeugt werden, indem die zulässigen Abweichungen in Tabellenfeldern schrittweise verkleinert werden. Die zulässigen Abweichungen ergeben sich einerseits aus dem Typ des Tabellenfeldes (ein- oder mehrdimensionale Auswertung) sowie andererseits aus der Größe des Wertes im Tabellenfeld, d. h. $f_i = b_i + g_i$. Da diese Werte sowohl von dem Typ der Auswertung und der Größe des Tabellenfeldes abhängen, werden sie als entsprechende zweidimensionale Liste behandelt (Tabelle 3). Bei mehrdimensionalen Tabellenfeldern wird eine um l erhöhte Abweichung erlaubt als bei den eindimensionalen.

Für die Korrektur der Lösung werden dem Algorithmus zwei Parameter übergeben, die die Zielparameter der durch die Korrektur zu erreichenden Abweichungsschranken für ein- und mehrdimensionale Tabellenfelder beschreiben. Das Programm versucht, beginnend bei den durch die Lösung der Aufgabe 1 erreichten Maximalabweichungen, durch schrittweises unabhängiges Verkleinern der einzelnen gewichteten bounds (Summe aus bound-Parameter und Gewicht) die Korrektur stufenweise durchzuführen. Es werden die in Tabelle 3 erhaltenen gewichteten bounds von links oben beginnend um 1 reduziert, bis eine weitere Reduzierung die Anzahl der so entstehenden Schrankenverletzungen eine vorgegebene Maximalzahl übersteigen würde. Beim Reduzieren wird die Regel beachtet, dass gewichtete bounds für ein höheres Tabellengewicht nicht kleiner sein dürfen als für ein kleineres Tabellengewicht (siehe folgendes Beispiel). Danach wird der Datenbestand wie bei Aufgabe 1 iterativ bearbeitet und durch die Änderung der Häufigkeiten im Bereich der anonymen Lösungen diese Anzahl verkleinert. Nicht anonyme Häufigkeiten sind im Korrekturlauf nicht mehr zulässig. Ist die Anzahl der bounds-Verletzungen 0, werden die bounds für die Tabellenfelder neu berechnet, die noch nicht die Zielparameter erreicht haben, und die Aufgabe 2 neu gelöst. Sind die Zielbounds erreicht oder beim weiteren Verkleinern eines einzelnen bounds wird die Anzahl der Schrankenverletzungen so groß, dass die Lösbarkeit als unwahrscheinlich eingeschätzt wird, wird das Programm beendet.

3 | Abweichungen nach Größenklassen und Dimensionen der Tabellenfelder

Typ der Auswertung	Gewicht g_i des Tabellenfeldes t_i					
	0	1	2	3	4	5
Eindimensional	k	k+1	k+2	k+3	k+4	k+5
Mehrdimensional	k+l	k+l+1	k+l+2	k+l+3	k+l+4	k+l+5

c | Programmablauf „Optimierung der Lösung“



Beispiel:
Die Abweichung der Tabellenfehler sei das Ergebnis der Anonymisierung. Als Ergebnis der Korrektur wird eine Maximalabweichung von 8 angestrebt:

Abweichung	Gewichtung des Tabellenfeldes g_i								
	0	1	2	3	4	5	6	7	8
Ergebnis der Anonymisierung									
0	82 527	20 772	18 737	10 368	8 654	18 169	14 005	1 885	86
1	267 988	40 163	35 643	19 349	15 911	34 570	26 519	3 417	180
2	143 137	28 356	27 313	16 814	13 937	29 883	23 773	3 115	196
3	31 107	16 021	17 078	11 028	11 468	24 385	19 265	2 524	141
4	6 625	5 827	11 621	6 926	6 592	19 210	14 501	1 932	122
5	522	1 550	3 254	5 302	4 118	8 986	10 208	1 405	116
6	15	154	885	1 150	3 521	5 260	4 255	992	90
7	0	11	88	265	527	5 413	2 310	480	67
8	0	0	4	25	102	431	2 609	274	30
9	0	0	0	0	7	58	84	251	17
10	0	0	0	0	0	3	14	23	20
11	0	0	0	0	0	1	1	1	2
12	0	0	0	0	0	0	1	0	1
13	0	0	0	0	0	0	0	0	1
14	0	0	0	0	0	0	0	0	0
>= 15	0	0	0	0	0	0	0	0	0
Maximal-abweichung	6	7	8	8	9	11	12	11	13
Parameter der bounds	6	7	8	8	8	10	11	11	12
Ergebnis des 1. Durchlaufs der Korrektur									
0	82 238	20 660	18 684	10 302	8 512	18 317	14 071	1 866	95
1	267 800	40 168	35 493	19 453	16 102	34 613	26 840	3 599	196
2	143 012	28 319	27 301	16 724	14 142	30 242	24 100	3 175	195
3	31 602	15 878	17 591	11 179	11 190	24 469	19 378	2 611	164
4	6 700	6 030	11 215	7 305	7 217	18 019	13 906	1 900	103
5	554	1 602	3 313	4 783	4 567	9 787	9 286	1 218	111
6	15	186	964	1 153	2 514	6 599	5 230	828	70
7	0	11	61	318	572	3 688	3 368	590	70
8	0	0	1	10	21	539	1 189	389	45
9	0	0	0	0	0	96	155	101	18
10	0	0	0	0	0	0	22	19	2
11	0	0	0	0	0	0	0	3	0
>= 12	0	0	0	0	0	0	0	0	0
Maximal-abweichung	6	7	8	8	8	9	10	11	11
Parameter der bounds	6	7	8	8	8	8	9	10	10
Ergebnis des 17. Korrekturdurchlaufs zur Parameterneubestimmung									
0	81 778	20 212	18 706	10 030	8 368	18 783	15 665	2 099	150
1	267 313	39 945	34 415	19 276	16 048	35 275	28 726	4 030	215
2	142 585	28 529	27 365	16 393	13 926	30 858	24 709	3 387	201
3	32 530	15 818	18 164	11 463	11 010	24 447	19 414	2 676	172
4	7 041	6 317	10 788	7 491	75 22	18 163	14 861	1 949	107
5	654	1 802	3 852	4 564	4 821	10 813	8 440	11 61	91
6	20	225	1 242	1 517	2 259	5 170	3 765	595	74
7	0	6	90	464	827	2 670	1 842	371	53
8	0	0	1	29	56	190	123	31	6
>= 9	0	0	0	0	0	0	0	0	0

Die Tabellenfelder, die die gesetzten bounds überschreiten, sind grau hinterlegt.

Nach 17 Korrekturdurchläufen (mit bounds-Parameterbestimmungen) ist das endgültige Korrekturergebnis erreicht.

Es erfolgt der Abbruch der Korrektur, da das Ziel einer Maximalabweichung von 8 erreicht ist.

Innerhalb der Aufgabe 2 wird eine neue Straffunktion für die Lösungskorrektur verwendet (S_K). Diese Straffunktion versucht Tabellenfelder mit der gerade zulässigen Maximalabweichung verstärkt zu verändern. Diese Abweichungen erhalten einen deutlich erhöhten Strafterm in der neuen Straffunktion. Es gilt wieder die gleiche Abstufung zwischen den Zielfunktionen. Die erste Zielfunktion sind die gesetzten Abweichungsschranken. Es werden wieder Veränderungskombinationen in Satzgruppen getestet. Werden dabei Sätze verändert, die in einem Tabellenfeld die bounds überschritten haben, so sind Veränderungen nur möglich, wenn sie diese Überschreitung danach beseitigen (ZF2). Dadurch wird verhindert, dass eine Überschreitung von ±1 durch die Änderung vergrößert wird. Danach werden die Veränderungen an der Veränderung in ZF3 und bei dort neutralen Veränderungen an ZF4 ausgewählt:

$$S_K = \sum_{i=1}^{k} s_{K\,i}$$

$$s_{Ki} = \begin{cases} 500 & ;\forall\left(\left|t_i^z - t_i^o\right| = f_i\right) \\ 9 & ;\forall\left(\left|t_i^z - t_i^o\right| = f_i - 1\right) \\ 4 & ;\forall\left(\left|t_i^z - t_i^o\right| = f_i - 2\right) \\ 1 & ;\forall\left(\left|t_i^z - t_i^o\right| = f_i - 3\right) \\ 0 & ;\forall\left(\left|t_i^z - t_i^o\right| < f_i - 3\right) \end{cases}$$

8. Rematching – Zuordnung zu Identifikatoren

Als separater Programmbaustein steht eine Routine zum Rematching zur Verfügung. Ergebnis des SAFE-Laufs ist eine Datei, in der die Merkmalskombinationen mit ihrer originalen und anonymen Häufigkeit stehen.

Das Programm nimmt eine Zuordnung der mit der SAFE-Anonymisierung erhaltenen anonymen Lösung zu den Sätzen des originalen Datenbestandes vor. Damit kann auch ein Zurückspielen von anonymen Daten anstelle der originalen in Auswertungsdatenbanken erfolgen. Diese müssen dann nicht als unabhängige Daten ausgewertet werden, sondern können auch die DB-Verknüpfungen zu anderen Tabellen oder quantitativen Werten weiterhin verwenden.

Dazu wird für die Merkmale des Datenbestandes eine Prioritätsreihenfolge festgelegt, die angibt, welche Merkmale/Merkmalsstufen wann aus dem Vergleich ausgeschlossen werden sollen. Die Anzahl der Merkmalsstufen sei *m*.

Für die Zuordnung wird folgender Algorithmus abgearbeitet:

1. Sortiere den originalen und den anonymen Datenbestand in der gewählten Prioritätsreihenfolge der Merkmale/Merkmalsstufen.
2. Setze die Anzahl der aktuell in den Vergleich einzubeziehenden Merkmale i auf i = m.
3. Gruppiere den noch nicht zugeordneten Datenbestand der anonymen Lösung in Gruppen mit identischen Ausprägungen in den ersten i-Merkmalen. Gruppiere den originalen Datenbestand analog. Wähle die erste Gruppe des Datenbestandes aus.
4. Für die gewählte Gruppe des anonymen Datenbestandes (Menge A) suche die analoge Gruppe aus dem noch nicht zugeordneten originalen Datenbestand (Menge B), die in den ersten i-Merkmalen die gleichen Ausprägungen hat.
5. Solange die Menge B nicht leer ist, suche für den ersten Satz der Menge A aus der Menge B den Satz heraus, der in den meisten Merkmalen (auch den nicht mehr in die Gruppierung einbezogenen) mit dem gewählten Satz aus A übereinstimmt. Kennzeichne diese beiden Sätze als Zuordnungspaar und entferne den Satz aus der Menge A (der „noch nicht zugeordneten" Sätze) und den Partner der aus der Menge B der originalen nicht zugeordneten Sätze (Kennzeichnung als nicht mehr „für Zuordnungen verfügbar"). Sind die Mengen A und B nicht leer, bearbeite die verbleibenden Sätze der Menge A durch Wiederholung des Schrittes 4.
6. Ist die Menge B leer (keine möglichen Partner vorhanden) oder die Menge A komplett zugeordnet, so bearbeite die nächste Gruppe des anonymen Datenbestandes (neue Menge A). Dazu wird diese Gruppe wieder ab Schritt 4 bearbeitet. Ist keine nächste Gruppe im anonymen Datenbestand vorhanden, so gehe zu Schritt 7.
7. Wenn alle Sätze zugeordnet sind (spätestens bei Durchlauf mit i = 0), dann gehe zu Schritt 8. Wenn noch nicht alle Sätze zugeordnet sind, so reduziere die zum Vergleich einzubeziehenden Merkmale um 1 (i = i – 1) und beginne wieder mit Schritt 3.
8. Exportiere alle gespeicherten Satzpaare in dem gewählten Speicherformat als Ausgabedatei.

9. Realisierung

Programmtechnischer Schwerpunkt des Verfahrens ist ein schnelles Testen der einzelnen Satzgruppen mit ihren möglichen Veränderungen und deren Auswirkungen auf die Randsummentabellen. Das erfordert, dass alle Randsummentabellen verfügbar sind und zu jedem Satz der Basisdatei schnell alle zugehörigen Tabellenfelder und die aktuellen Fehler gelesen werden können. In der Aufgabenstellung ist es zwar eine klassische Datenbankanwendung, da die Abfragen sich für jeden Satz einer Satzgruppe und jede Tabelle unterscheiden, entstehen jedoch sehr viele unterschiedliche Abfragen, die mit einer Datenbankanwendung nicht performant realisiert werden können. Deshalb wurde das Programm in C geschrieben. Um die Suchalgorithmen zu beschleunigen, wurde ein spezieller Indextyp entwickelt. Dieser ähnelt einem graphischen Index, wie er für hierarchische Datenbanken verwendet wird. Er hat neben dem schnelleren Zugriff auch den Vorteil, dass er weniger Speicherplatz erfordert und somit auch größere Probleme vollständig im Hauptspeicher realisierbar sind. Das Programm wird inzwischen als 32- oder 64-Bit-Mehrprozessor-Anwendung gewartet und weiterentwickelt.

10. Sicherheit und Qualität des Verfahrens

a) Sicherheit

Die Datensicherheit lässt sich beim Verfahren SAFE klar beurteilen. SAFE ist ein datenveränderndes Verfahren. Die Sicherheit entsteht hier nicht durch das Unterdrücken oder Löschen von sensiblen Informationen, sondern durch die Unsicherheit, ob Informationen verändert wurden. Jede Kombination an Merkmalsausprägungen ist dreifach vorhanden, wobei dies durch die Veränderung von exotischen Merkmalskombinationen entstanden sein kann. Das Nichtvorhandensein von Merkmalskombinationen bedeutet nicht, dass es im Originaldatenbestand diese Kombination nicht gegeben hat. Randwerte in Tabellen können daher nicht mehr als Angriffswissen verwendet werden. Die Information, dass alle Merkmalsträger bei einer Variablen die gleiche Ausprägung aufweisen, kann auch durch die Veränderungen in SAFE entstanden sein. Ein „Datenangreifer" kann daher nicht mit Sicherheit ableiten, dass alle Merkmalsträger die ausgewiesene Ausprägung gemeldet haben.

Das Verfahren hat den Vorteil, dass der Datenbestand nur einmal anonymisiert werden muss und dann alle Auswertungen aus diesem anonymen Datenbestand erzeugt werden können. Da der anonyme Datenbestand eine 3-anonyme Datei erzeugt hat, sind alle Auswertungstabellen sicher. Eine manuelle Prüfung jeder einzelnen Auswertungstabelle auf die Einhaltung des Statistikgeheimnisses entfällt. Kleine Fallzahlen werden in keiner Tabelle mehr ausgewiesen.

b) Qualität bei Anwendung als Tabellengeheimhaltungsverfahren

Das SAFE-Verfahren erzeugt bei jedem Durchlauf einige Kennzahlen zur Qualität des Anonymisierungslaufs. Mit diesen Kennzahlen kann die Qualität des Tabellengeheimhaltungsverfahrens beurteilt werden. Es können auch verschiedene Varianten, die sich beispielweise durch unterschiedliche Kontrolltabellensets auszeichnen, verglichen werden. Es werden Häufigkeitstabellen erstellt, wie oft welche absoluten Abweichungen sowohl in ein- als auch in zwei- und höherdimensionalen Tabellenfeldern je nach Größenklasse auftreten. In diesen Dateien kann der größte Fehler je Größenklasse abgelesen werden. Die Häufigkeitstabellen können als Ausgangspunkt genommen werden, um durchschnittliche Abweichungen in eindimensionalen sowie in zwei- und höherdimensionalen Tabellenfeldern zu berechnen. Auch können kumulierte Angaben gemacht werden, im Stil von „in 90 % aller Tabellenfelder ist die Abweichung kleiner oder gleich 3". Qualitätskennzahlen in dieser Art werden zu den Testrechnungen mit den Einzeldaten der Volkszählung 1987 in Westdeutschland in Höhne (2011) berichtet.

Diese Qualitätsangaben werden nur für Kontrolltabellen berechnet. Nicht-Kontrolltabellen können, wenn die Merkmale im Anonymisierungslauf enthalten waren, aus dem anonymen Datenbestand auch flexibel erzeugt werden und sind ebenfalls sicher. Eine Aussage über deren Qualität ist ex ante aber nicht möglich.

Die Interpretation von mit datenverändernden Verfahren geheim gehaltenen Tabellen unterscheidet sich von anderen Tabellenergebnissen. Der Informationsverlust entsteht, wie oben bereits beschrieben, nicht durch die Unterdrückung von Informationen, sondern durch Unsicherheit in den Informationen aufgrund der Veränderung. Dabei sind Veränderungen um ±2 nötig. Betrachtet man die relativen Veränderungen von Tabellenfeldern, so sind diese bei kleinen Tabellenfeldern besonders groß. Gerade bei den kleinen Tabellenfeldern, hier insbesondere die mit den Häufigkeiten 1 und 2, besteht aber der Schutzbedarf.

Das Verfahren löst aufgrund der Mikroaggregation das Problem, dass keine zu kleinen Fallzahlen mehr auftreten werden. Das Randsummenproblem ist aufgrund der Unsicherheit bei der Interpretation auch nicht mehr vorhanden.

11. Anwendung bei Dateien mit verschiedenen statistischen Einheiten

In einigen Bundesstatistiken sind Angaben zu verschiedenen statistischen Einheiten gemeinsam in einem Mikrodatenbestand enthalten oder die Angaben aus zwei Erhebungen oder Satzarten können verknüpft und kombiniert ausgewertet werden. Solche Datenbestände mit einer hierarchischen Struktur können mit dem SAFE-Verfahren ebenfalls geheim gehalten werden. Dabei sind jedoch einige Aspekte zu beachten. Diese sollen beispielhaft anhand einer fiktiven Pflegestatistik beschrieben werden.

In der Pflegestatistik sind u. a. Angaben zu den Einrichtungen und den Pflegebedürftigen enthalten. Die Einzeldaten beider statistischen Einheiten sollen zu einem Datensatz verknüpft werden, sodass alle Merkmale als Eigenschaften der kleinsten Ebene (pflegebedürftige Personen) gespeichert werden. Bei der Pflegestatistik sollen die Angaben zu den Einrichtungen mit den Angaben der Pflegebedürftigen über die Einrichtungsnummer verknüpft werden. In jeder Einrichtung werden dabei mehrere Pflegebedürftige betreut. Der neue Datenbestand weist in jeder Zeile die Angaben zu einem Pflegebedürftigen auf. Deshalb kann man die Pflegebedürftigen als die führende statistische Einheit in diesem neuen Datenbestand bezeichnen. Die Angaben jeder Pflegeeinrichtung sind nun mehrfach im Datenbestand enthalten, da sie bei jedem Pflegebedürftigen dieser Einrichtung angespielt wurden. Bei Auswertungen der Pflegeeinrichtungen kann dieser

4 | Beispieldatensatz mit verschiedenen statistischen Einheiten

Pflegeeinrichtung					Pflegebedürftige		
ID_Inst	Plätze	An-gestellte	Träger	Zähler_ Einr	ID_ Person	Alter	Pflege-stufe
25	210	120	2	1	233	70	2
25	210	120	2	0	234	85	3
25	210	120	2	0	236	84	2
27	67	35	3	1	238	84	3
27	67	35	3	0	239	81	3

Datenbestand nicht direkt ausgewertet werden, da Mehrfachzählungen der Einrichtungen auftreten würden. Vielmehr muss eine Filtervariable (Zähler) eingefügt werden, die gewährleistet, dass jede Einrichtung nur genau einmal gezählt wird. Bei allen Auswertungen nach Pflegeeinrichtungen wird die Auswertung mit dieser Filtervariable gekreuzt. Sollen die verschiedenen Träger ausgewertet werden, so muss die Häufigkeitstabelle Träger mit der Filterbedingung Zähler_Einr = 1 berechnet werden.

Wendet man das SAFE-Verfahren auf diesen Datenbestand mit verschiedenen statistischen Einheiten an, so entsteht wiederum ein 3-anonymer Datenbestand, aus dem alle Auswertungen geheim sind. Jede Merkmalskombination ist entweder mindestens dreifach oder nicht vorhanden. Dies gilt sowohl für Auswertungen der statistischen Einheit „Pflegebedürftige" als auch der Einheit „Pflegeeinrichtungen", wobei diese auch beim Auswerten des anonymen Bestandes stets mit dem Zähler-Feld kombiniert sein müssen.

Wird bei der Anonymisierung gewährleistet, dass die Kreuzkombinationen mit dem Einrichtungszähl-Feld (Zähler_Einr) als Kontrolltabellen berücksichtigt werden, so gelten die Qualitätsaussagen sowohl für die Auswertungen nach Pflegebedürftigen als auch nach Einrichtungen.

Die Merkmale „Anzahl der Plätze" und „Anzahl der Angestellten" könnten für die Geheimhaltung auch in Größenklassen klassiert werden.

12. Zusammenfassung

Mit SAFE steht ein Verfahren bereit, das einerseits Einzeldaten so anonymisiert, dass man einen anonymen Datenbestand freigeben kann, andererseits ist es ein pre-tabulares Geheimhaltungsverfahren, sodass keine Geheimhaltungsfälle in Auswertungstabellen mehr enthalten sind, die aus dem anonymen Datenbestand erzeugt werden. Das Verfahren optimiert die Lösung für ein vorgegebenes Set an Kontrolltabellen, für die das Verfahren direkt Qualitätsmaße angibt. Gleichzeitig bleibt die flexible Auswertbarkeit der Einzeldaten gewährleistet.

Dr. Jörg Höhne leitet die Abteilung *Gesamtwirtschaft* im Amt für Statistik Berlin-Brandenburg. Er studierte Statistik und Wirtschaftsmathematik in Berlin und Moskau und promovierte 2009 an der Universität Tübingen mit einer Arbeit über „Verfahren zur Anonymisierung von Einzeldaten".

Literatur

- Appel, Günther; Kinzel, Sabine; Nölte, Dieter (1993): SAFE – A Generally Usable Program System for the Anonymization of Individual Data in Official Statistics. In: Proceedings of the International Seminar on Statistical Confidentiality, Dublin, Ireland, 8–10 September 1992, S. 201–228.
- Höhne, Jörg (2003): SAFE – Ein Verfahren zur Anonymisierung und statistischen Geheimhaltung wirtschaftsstatistischer Einzelangaben. In: Berliner Statistik, Statistische Monatszeitschrift, Nr. 3/2003, Berlin 2003, S. 96–107.
- Höhne, Jörg (2008): Anonymisierungsverfahren für Paneldaten. In: Springer, Wirtschafts- und Sozialstatistisches Archiv Band 2/2008, S. 259–275.
- Höhne, Jörg (2010): Verfahren zur Anonymisierung von Einzeldaten. Statistik und Wissenschaft Band 16, Statistisches Bundesamt, Wiesbaden.
- Höhne, Jörg (2011): SAFE – A method for anonymising the German Census. Working Paper 16 at the Joint UNECE/ Eurostat work session on statistical data confidentiality, 26–28 October 2011, Tarragona. http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ ge.46/2011/16_Germany.pdf
- Höhne, Jörg (2012): Statistische Geheimhaltung des Zensus 2011. Vortrag im Rahmen der Statistik-Tage Bamberg Fürth 2012, „Die Methoden und Potenziale des Zensus 2011" am 26. und 27. Juli 2012. https://www.statistik.bayern.de/medien/wichtigethemen/st_vortrag_hoehne_27072012.pdf
- Hundepool, Anco; Domingo-Ferrer, Josep; Franconi, Luisa; Giessing, Sarah; Schulte Nordholt, Eric; Spicer, Keith; de Wolf, Peter-Paul (2012): Statistical Disclosure Control. Wiley Series in survey methodology, John Wiley & Sons Ltd, Chichester. Überarbeitete Fassung des ESSNET-SDC Handbook. http://neon.vb.cbs.nl/casc/SDC_Handbook.pdf
- Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik (Hrsg.): Wege zu einer besseren informationellen Infrastruktur. Gutachten der vom Bundesministerium für Bildung und Forschung eingesetzten Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik Baden-Baden 2001.
- Lechner, Sandra; Pohlmeier, Wilfried (2003): Schätzung ökonometrischer Modelle auf der Grundlage anonymisierter Daten. In: Gnoss, Roland; Ronning, Gert (Hrsg.): Anonymisierung wirtschaftsstatistischer Einzeldaten, Schriftenreihe „Forum der Bundesstatistik", Band 42, S. 115–137, hrsg. vom Statistischen Bundesamt, Wiesbaden.
- Lenz, Rainer (2010): Methoden der Geheimhaltung wirtschaftsstatistischer Einzeldaten und ihre Schutzwirkung. Statistik und Wissenschaft Band 18, Statistisches Bundesamt, Wiesbaden.
- Ronning, Gerd; Sturm, Ronald; Höhne, Jörg; Lenz, Rainer; Rosemann, Martin; Scheffler, Michael; Vorgrimler, Daniel (2005): Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten. In: Schriftenreihe „Statistik und Wissenschaft", Band 4, hrsg. vom Statistischen Bundesamt, Wiesbaden.
- Sweeney, Latanya (2002): k-anonymity – a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; S. 557–570.
- Zühlke, Sylvia; Christians, Helga; Cramer, Katharina (2007): Das Forschungsdatenzentrum der Statistischen Landesämter – eine Serviceeinrichtung für die Wissenschaft. AStA Wirtschafts- und Sozialstatistisches Archiv 3-4, S. 169–178.