Jörg Höhne

SAFE – ein Verfahren zur Geheimhaltung und Anonymisierung statistischer Einzelangaben

1. Einleitung

Die amtliche Statistik erhebt Einzeldaten von Personen, wirtschaftlichen Einheiten und anderen Merkmalsträgern und generiert aus diesen Einzeldaten Ergebnisse in vielfältiger Gliederung für verschiedenste Verwendungszwecke. Die Einzeldaten bleiben dem Nutzer der Statistik daher verborgen. Dieses ist aufgrund der Geheimhaltungsverpflichtung erforderlich, bedeutet jedoch Informationsverlust und hat zur Folge, dass kaum weitere Auswertungen möglich sind.

Um die Geheimhaltung zu gewährleisten, werden in statistischen Ergebnissen die Einzelangaben nach bestimmten Kriterien, z. B. Wirtschaftszweigen, Betriebsgrößenklassen, Altersgruppen, Regionen gegliedert und zusammengefaßt und noch vorhandene Einzelangaben unterdrückt oder durch weitere Vergröberung von Gliederungen unsichtbar gemacht.

Ein anderer Weg der statistischen Geheimhaltung besteht darin, sicherzustellen, dass die Einzeldaten nicht mehr ihren Merkmalsträgern zugeordnet werden können. Das erfolgt beispielsweise durch das gezielte Verändern einzelner Merkmale. Damit bleibt das Informationspotential der Einzeldaten für Verlaufsanalysen u. a. dem Nutzer erhalten.

Die Suche nach Verfahren zur Anonymisierung statistischer Einzeldaten hat vor allem nach Inkrafttreten des Bundesstatistikgesetzes von 1987 [1] neuen Auftrieb erhalten. Das Gesetz ermöglicht es, der Wissenschaft faktisch anonymisierte Einzeldaten – Scientific Use Files – zur Verfügung zu stellen. Der Bedarf der Wissenschaft wird nicht zuletzt durch die jüngsten Empfehlungen der Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik dokumentiert [2].

Die hier beschriebene Methode der Geheimhaltung ist ein Verfahren der Mikroaggregation. Es ist das Ergebnis längerer Testreihen und wurde an den Daten des „Monatsberichts für Betriebe" und der „Jahreserhebung für Betriebe des Verarbeitenden Gewerbes" für Berlin einer ersten Evaluation unterzogen. Der Teil der qualitativen Geheimhaltung wurde außerdem am Einwohnerregister Berlins mehrfach getestet. Das Einwohnerregister stellt durch seine Größenordnung (ca. 3,5 Mio. Sätze) eine Herausforderung für die numerischen Verfahren dar.

Diskussionswürdig für die weitere Ausgestaltung des Verfahrens ist das Ziel seiner Anwendung. Die bisherigen Untersuchungen erfolgten zur Schaffung eines Datenkörpers aus Einzelangaben, aus dem sofort konsistent alle potenziellen Auswertungen in anonymisierter Form erstellt werden können. Die Gesamtstruktur und ihre Zusammenhänge werden dabei gut reproduziert. Sollen die Einzeldaten beispielsweise für wissenschaftliche Untersuchungen genutzt werden, sind weitere Optimierungen in den angewendeten einzelnen Schritten und ihren Zielfunktionen durchaus möglich und sinnvoll.

Die Mikroaggregation orientiert sich ständig an einem vorher festgelegten Satz von Auswertungstabellen, der mit der anonymen Basisdatei erzeugt werden soll. An diesem Satz von Tabellen werden alle Geheimhaltungsfälle markiert und für diese Zellen Unzulässigkeitsbereiche bestimmt.

Danach erfolgt die Anonymisierung in folgenden Schritten:

1. Lösung der qualitativen Geheimhaltung,
2. Zuordnung der Lösung,
3. Lösung der quantitativen Geheimhaltung,
4. Optimierung der Lösung,
5. Gruppierung und Durchschnittsbildung.

2. Begriffsbestimmungen

Eine *Basisdatei* ist eine Datei, in der jedes statistische Objekt (Merkmalsträger) durch einen einzelnen Datensatz (Zeile) repräsentiert wird. Sie bildet den Ausgangspunkt für alle möglichen Auswertungen des Datenbestandes. Die Merkmale in den Basisdateien unterteilen sich in die zwei Kategorien „qualitative Merkmale" und „quantitative Merkmale". Die, für die Geheimhaltung betrachteten, Basisdateien enthalten keine Identifikationsmerkmale.

Bei *qualitativen Merkmalen* handelt es sich um Merkmale, die eine diskrete, feste Anzahl an Ausprägungen haben. Die Ausprägungen sind in einer Schlüsseltabelle zusammengefasst. Qualitative Merkmale können durch Umschlüsseln auf andere Schlüsselebenen umgesetzt werden. Beispielsweise können Regionalschlüssel wie Adresse aus Straße und Hausnummer auch in Statistisches Gebiet, Gemeinde, Land usw. umgeschlüsselt werden. Analog kann der Wirtschaftszweig in Branchen oder die einzelne Nationalität in Deutsch/Nichtdeutsch umgeschlüsselt werden. *Schlüsseltabellen* sind die Vorschriften, mit denen eine Umschlüsselung vorgenommen werden kann. In jeder Schlüsselstufe sind die Ausprägungen so gebildet, dass sie die Gesamtheit des Datenbestandes beschreiben können. Innerhalb der Basisdatei wird jedes der Merkmale in seiner feinsten Ausprägung gespeichert.

Quantitative Merkmale sind in der amtlichen Statistik, bedingt durch die Messgenauigkeit, meist ganzzahlig. Es gibt keine endliche vorher festgelegte Schlüsselmenge, die zulässig ist (Beispiele: Umsatz, Beschäftigte).

Aus quantitativen Merkmalen lassen sich durch Gruppenbildung wieder qualitative Klassen erzeugen, (z. B. Betriebe mit unter 20 Beschäftigte, 20 bis unter 50 Beschäftigte usw.).

Identifikationsmerkmale (Ident-Nummern, Betriebsnummern usw.) sind Schlüsselmerkmale, die eine eindeutige Zuordnung des Datensatzes zum statistischen Objekt ermöglichen. Diese müssen bei anonymisierten Dateien in jedem Fall entfernt sein.

Eine Basisdatei B ist eine Menge aus n-Objekten. Jedes Objekt wird durch ein Tupel von l+k Werten beschrieben. Dabei beschreiben die Werte $b_{i,1},\ldots,b_{i,l}$ die qualitativen Merkmale des i-ten Objektes und die Werte $b_{i,l+1},\ldots,b_{i,l+k}$ die quantitativen. Die Reihenfolge der Merkmale, zuerst qualitative, dann quantitative, ist durch einfaches Um-

sortieren jederzeit herstellbar. Existieren keine quantitativen Merkmale, so handelt es sich um einen Spezialfall der Anonymisierung von nur qualitativen Feldern wie z. B. beim Datenbestand des Einwohnerregisters.

Beispiel:
B - Basisdatei mit $B=\{b_{ij}\}$ – i=1,2,…,n
– j=1,2,…,l,l+1,…,l+k
n – Anzahl der statistischen Objekte
l – Anzahl der qualitativen Merkmale
k – Anzahl der quantitativen Merkmale
b_{ij} – Wert des Merkmals j beim Objekt i

Neben der Basisdatei existieren zu jedem qualitativen Merkmal Schlüsseltabellen, mit deren Hilfe das Merkmal von einer Stufe in eine andere Stufe umgeschlüsselt werden kann. Die Schlüsseltabellen sollten für alle üblichen Merkmalsaggregationen bereitgestellt werden.

Für die Realisierung des Verfahrens wurden hierarchische Schlüsselstufen unterstellt. Damit können für jeden Schlüssel Nummerierungen in der unten dargestellten Art erzeugt werden (integrierte Schlüssel).

Beispiel integrierter Schlüssel (Nationalitäten):

Länder		**Regionen**		**Deutsch / Nichtdeutsch**	
1101	Deutschland	11	Deutschland	1	Deutsch
2101	Frankreich	21	Europa	2	Ausländer
2102	Polen				
…	usw.				
2201	Argentinien	22	Amerika		
…	…				

Die Schlüssel stimmen nicht unbedingt mit den amtlichen Schlüsseln überein. Beispiele für amtliche hierarchische Schlüssel wären u. a. die Systematik der Wirtschaftszweige (WZ93) und die amtlichen Regionalschlüssel. Der Wechsel zwischen den verschiedenen Stufen von Schlüsseln ist durch einfaches Abschneiden des Schlüssels möglich. Integrierte Schlüssel werden zwar bei der derzeitigen programmtechnischen Lösung unterstellt, sind jedoch nicht zwingend für das Verfahren erforderlich. Für jedes qualitative Merkmal ist somit die Anzahl der möglichen Schlüsseltabellen und die dazu erforderliche Schlüssellänge bekannt.

Für jedes quantitative Merkmal existieren ebenfalls Vorschriften zur Gruppenbildung. Diese stellen sich aber in Form von Intervallgrenzen dar:
Beispiel:
Betriebsgrößen nach Beschäftigtengrößenklassen

	unter	5	Beschäftigte
5	bis unter	10	Beschäftigte
10	bis unter	20	Beschäftigte
20	bis unter	50	Beschäftigte
50	bis unter	100	Beschäftigte
100	bis unter	200	Beschäftigte usw. … .

Für ein Merkmal sind auch verschiedene Gruppengrenzen möglich. Mit Hilfe dieser Intervallgrenzen lässt sich die Basisdatei um zusätzliche qualitative Felder erweitern.

Alle üblichen Gruppenbildungen müssen für jedes Merkmal als Information bekannt und gegeben sein.

Aus den obigen Informationen besteht die Möglichkeit, alle potenziellen Auswertungsmöglichkeiten für die Basisdatei zu generieren. Dabei werden unter den Auswertungsmöglichkeiten alle Kombinationen aus einem und mehr (meist bis 3) qualitativen Merkmalen mit beliebigen Schlüsselstufen verstanden. Für diese Kombinationen erfolgt eine Aggregation der Basisdatei durch Gruppenbildung (Aufsummierung von Sätzen mit gleicher Ausprägung).

Für eine Auswahl aus einem oder mehr qualitativen Merkmalen in einer beliebigen Schlüsselstufe wird eine komplette Auflistung aller vorhandenen Ausprägungskombinationen und die Aggregation der quantitativen Werte der Basisdatei als Auswertungstabelle bezeichnet. Eine feste Kombination von Merkmalsausprägungen beschreibt einen Tabellenwert.

Alle Auswertungstabellen sind somit eine Abbildung aus der Basisdatei:

T =F(B) – Vektor der Ergebnisse aller Auswertungstabellen (Der Vektor entsteht durch Aneinanderreihung aller Tabellenwerte über alle möglichen Auswertungstabellen)
$t_{p,q,m}=f_{p,q,m}(B)$ – Wert des Merkmals m bei der Ausprägungskombination q (q-te Tabellenwert) in der Auswertungstabelle p

3. Geheimhaltungsprobleme

Aus §16 Bundesstatistikgesetz (BStatG) [1]ergibt sich die Verpflichtung, „Einzelangaben über persönliche und sachliche Verhältnisse … geheimzuhalten".

Für jedes generierbare Tabellenfeld kann ein Geheimhaltungsproblem auftreten. Das ist dann der Fall, wenn aus den Tabellenfeldern Rückschlüsse auf ein einzelnes statistisches Objekt (Unternehmen, Bürger usw.) gezogen werden können, und man so Informationen über das statistische Objekt erhält, die nur aufgrund der statistischen Veröffentlichung möglich sind.

Folgende Geheimhaltungsprobleme können bei der Tabellierung entstehen:

a) Fallzahlprobleme

Fallzahlprobleme treten auf, wenn der Inhalt eines Tabellenfeldes nur aus dem Merkmalswert eines oder zweier Objekte entstanden ist. Bei einem Einer-Fallzahlproblem ist der Tabellenwert mit dem Merkmalswert eines Unternehmens identisch, so dass hier ein Geheimhaltungsbedarf besteht. Es wird davon ausgegangen, dass in diesem Fall die Identifikation des Merkmalsträgers durch die Definition des Tabellenfeldes nicht ausgeschlossen werden kann. So ist beispielsweise bezüglich der Leistung des öffentlichen Personennahverkehrs einer Großstadt oftmals allseits bekannt, dass es nur ein Unternehmen dieser Kategorie gibt. Bei zwei statistischen Objekten besteht das Risiko darin, dass das eine Objekt durch Differenzbildung die Information über das andere Objekt (in der Regel den Konkurrenten) problemlos generieren kann. Ein Fallzahlproblem besteht ebenfalls, wenn zwar einerseits mehr als zwei Objekte existieren, aber bekannt ist, dass das konkrete quantitative Merkmal nur bei einem oder zwei Objekten ausgeprägt ist (z. B. nur ein Unternehmen der Branche hat einen Auslandsumsatz).

b) Randsummenprobleme

Randsummenprobleme entstehen, wenn innerhalb einer Tabelle in einer Zeile oder Spalte nur eine Zelle belegt ist. In diesem Fall können auch mehr als zwei Objekte zur konkreten Ausprägung des Tabellenwertes beitragen. So beispielsweise in der Todesursachenstatistik: 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

inhaltsabhängig 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.

c) *Dominanzprobleme*

Dominanzprobleme entstehen, wenn innerhalb eines quantitativen Tabellenfeldes die Mengenangabe zu einem sehr großen Teil durch ein oder zwei Einzelobjekte beigetragen wird. In diesem Fall ist die Kenntnis der Gesamtmenge des quantitativen Feldes eine gute Schätzung des Einzelwertes des Objektes. Als Regeln für die Bestimmung des Auftretens von Dominanzproblemen existieren:

1-k-Regel

Hat das größte Objekt einen Anteil von mehr als k % an der Gesamtmenge des Tabellenfeldes (in der Regel 80 %), so wird die Gesamtmenge als zu gute Schätzung für den Einzelwert angesehen.

2-k-Regel

Die Summe der beiden größten Objekte übersteigt mehr als k % der Gesamtmenge des Tabellenfeldes (in der Regel 85 %), so wird die Gesamtmenge als zu gute Schätzung für den Einzelwert angesehen.

p %-Regel

Dominanz tritt auf, wenn das zweitgrößte Objekt durch Abzug seines Anteils von der Gesamtmenge eine Schätzung für das größte Objekt erhält, dessen Fehler weniger als p % (in der Regel 5 %) beträgt. Die Größe der Parameter p und k wird so gewählt, dass ein optimaler Kompromiß zwischen dem Geheimhaltungsbedarf und dem Informationsverlust durch Geheimhaltung entsteht. Dadurch können sich die Parameter abhängig von der konkreten Statistik variieren. Die p %-Regel bildet sowohl die Fallzahlprobleme als auch die Dominanzprobleme in einem gleichmäßig quantitativ fundierten Deanonymisierungsrisiko ab (siehe grafische Darstellung auf der Titelseite). In der amtlichen Statistik findet deshalb ein schrittweiser Übergang der Statistiken zur p %-Regel statt.

Für die grafische Darstellung von Geheimhaltungsregeln eignen sich folgende Darstellungen: Auf der X-Scala werde das größte und auf der Y-Scala das zweitgrößte Unternehmen dargestellt. Damit stellt sich jeder Tabellenwert als ein Punkt in den Diagrammen der Titelgrafik dar. Aus dem Zusammenhang zwischen dem größten und dem zweitgrößten Unternehmen ergeben sich folgende Beziehungen $Y \leq X$ und $X + Y \leq 100$. Damit liegen alle potenziell möglichen Tabellenwerte innerhalb des hell dargestellten Dreiecks. Die einzelnen Geheimhaltungsregeln der 1,k-Regel lassen sich innerhalb der Fläche folgendermaßen darstellen:

- Das „Einer-Fallzahlproblem" ist der rechte Eckpunkt des Dreiecks;
- das „Zweier-Fallzahlproblem" ist die rechte Kante des Dreiecks (hier gilt X+Y=100);
- das Dominanzproblem ist das markierte rechte Teilstück (bei der 1,80 %-Regel alles rechts des Wertes von 80 %).

Die drei Bereiche der geheimzuhaltenden Fälle (1,2 Fallzahlprobleme und Dominanzprobleme) werden bei der p %-Regel durch einen gleichmäßigen ineinander übergehenden Bereich gekennzeichnet, womit das sehr ungleiche Deanonymisierungsrisiko des größten Merkmalswertes bei Kenntnis des zweitgrößten aus der 1,k-Regel beseitigt wird.

Neben den Problemen aus der Darstellung bei Tabellen existieren bei Basisdateien noch Reidentifikationsprobleme, die sich aus so genannten „Matching"-Versuchen ergeben. Es wird versucht, Informationen, die man über einzelne Unternehmen aus externen Quellen gewonnen hat, gegen alle Sätze der Basisdatei gegenzuspielen und so bei einer eineindeutigen Übereinstimmung zusätzliche Eigenschaften aus der Basisdatei abzulesen. Es steht einem beispielsweise eine Basisdatei zur Verfügung, die Angaben über Wirtschaftszweig, Beschäftigte und Umsatz enthält. Hat man aus externen Quellen (wie Zeitungsberichten u.ä.) die Information über den Wirtschaftszweig und die aktuelle Anzahl der Beschäftigten eines Unternehmens erhalten, so liefert einem diese Basisdatei auch den Umsatz, wenn nur ein Unternehmen mit genau dieser Beschäftigtenanzahl im Wirtschaftszweig existiert. Diese Probleme sind den Fallzahlproblemen aus der Tabellengeheimhaltung sehr ähnlich. In der Basisdatei sind jedoch meist mehr Merkmale enthalten als in den Auswertungstabellen. Allen vier Deanonymisierungsrisiken wird im Rahmen des SAFE-Verfahrens Rechnung getragen.

4. Lösungsansatz SAFE

Das SAFE-Verfahren ist ein Verfahren der Mikroaggregation. Einzelne sich unterscheidende Datensätze einer Basisdatei werden durch gezielte Auswahl und Gruppenbildung so vereinheitlicht, dass jeder Datensatz in der Basisdatei mit mindestens zwei weiteren Sätzen in der Datei identisch ist.

Damit ergeben sich für die einzelnen Deanonymisierungsrisiken folgende Sicherheiten:

Fallzahlprobleme können in den Tabellen nicht mehr auftreten, da mindesten 3 Sätze zum Tabellenwert beitragen. Das bedeutet, dass entweder die in der Realität auftretenden Fallzahlprobleme entfernt wurden, oder durch die Aggregation die Häufigkeit der Ausprägungskombination auf mindestens 3 erhöht wurde.

Matching-Algorithmen können nur zu einer mehrdeutigen Zuordnung führen. Wenn ein Satz jedoch mehrere Entsprechungen in der anonymisierten Basisdatei hat, die wiederum durch Durchschnittsbildung und ggf. Gruppenbildung entstanden sind, so kann nicht daraus geschlossen werden, dass die zusätzlich ablesbaren Eigenschaften für das Original gelten.

Randsummenprobleme können zwar theoretisch bei Auswertungstabellen generiert werden, aber auch hier ist es durchaus möglich, dass diese Probleme nur das Ergebnis der Anonymisierungstechnik sind. Die Probleme können entstehen, da durch das Gruppieren Objekte mit qualitativ verschiedenen Eigenschaften, aber geringen Häufigkeiten, aus dem Datenbestand entfernt werden. Es ist somit kein sicherer Rückschluss auf die Eigenschaften der Basisdatei mehr möglich.

Dominanzprobleme sind die einzigen Geheimhaltungsprobleme, die nicht direkt durch die Mikroaggregation gelöst werden können. Hier ist erforderlich, bei der Bestimmung der Lösung Unzulässigkeitsbereiche zu definieren, die durch die anonymisierte Basisdatei nicht belegt werden dürfen. Aus diesem Grunde werden für die Geheimhaltung alle Auswertungstabellen und für die einzelnen Tabellenwerte mit Geheimhaltungsproblemen Unzulässigkeitsintervalle bestimmt, in denen die Lösung der gleichen Auswertungstabelle mit der anonymisierten Basisdatei nicht liegen darf. Solche Unzulässig-

keitsintervalle sind auch für die quantitativen Ausprägungen bei 1- und 2-Fallzahlproblemen zu bestimmen. Einen sehr guten methodischen Ansatz für die Bestimmung der Intervallgrenzen bietet die p %-Regel. Aber auch die 1-k-Regel lässt die Bildung eines solchen Intervalls zu, wenn man aus der Geheimhaltungspflicht von Tabellenwerten mit mehr als k %-Anteil (z. B. 80 %) am Gesamtwert den Umkehrschluss zieht, dass der veröffentlichbare Wert um mehr als 100*(100-k)/k vom größten Einzelwert abweicht (z. B. 100*(100-80)/80=25 % Fehler). Der erforderliche Datenfehler aus der 1-k-Regel ist jedoch bedeutend höher als bei der p %-Regel, was sich bei der zusätzlichen Anwendung auf alle Fallzahlprobleme bemerkbar macht.

Mit der oben eingeführten Notation lässt sich das Geheimhaltungsproblem folgendermaßen darstellen:

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

$T^o = F(B^o)$	– Vektor der Ergebnisse aller Auswertungstabellen
$t^o_{p,q,m} = f_{p,q,m}(B^o)$	– Tabellenwert des Merkmals m für die q-te Ausprägungskombination in der Auswertungstabelle p

Nach Bestimmung aller Geheimhaltungsfälle in den Auswertungstabellen lassen sich für die Geheimhaltungsfälle Unzulässigkeitsintervalle als untere Grenze ($z^u_{p,q,m}$) und obere Grenze ($z^o_{p,q,m}$) um den geheimzuhaltenden Wert bilden (s. o.).

Damit muss für eine anonymisierte Basisdatei B^a gelten:

$T^a = F(B^a)$ mit

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

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

$z^u_{p,q,m} = z^o_{p,q,m} = t^o_{p,q,m}$ wenn kein Geheimhaltungsfall bei Merkmal m in der Ausprägungskombination q der Tabelle p

T^a	– Vektor der Ergebnisse aller anonymen Auswertungstabellen
$t^a_{p,q,m} = f_{p,q,m}(B^a)$	– Tabellenwert für Merkmal m in der Ausprägungskombination q der anonymen Auswertungstabelle p

Neben diesen Randbedingungen muss in der Matrix B^a jede Zeile mit mindestens zwei weiteren Zeilen identisch sein. Gesucht ist natürlich 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 Abstandsfunktion hängt aber von den konkreten Bedürfnissen ab.

Die Problemkomplexität lässt sich u.a. wie folgt illustrieren:

- Die Aggregationsfunktionen zur Bestimmung der einzelnen Tabellenwerte sind bezüglich der qualitativen Merkmale nicht stetig.
- Die quantitativen Merkmale haben bei Geheimhaltungsfällen eine zulässige Lösungsmenge, die oberhalb und unterhalb eines unzulässigen Bereiches liegt. Die Lösungsmenge ist damit nicht konvex.
- Die Basisdatei kann mehrere Tausend statistische Objekte umfassen (Anzahl der Zeilen von B).

Die Lösung lässt sich deshalb nicht einfach mit klassischen Verfahren beispielsweise der linearen oder nichtlinearen Optimierung bestimmen. Es wird deshalb eine pragmatische aber effiziente Herangehensweise gesucht. Ausgangspunkt ist die Teilung des Problems in die getrennte Lösung der qualitativen und quantitativen Geheimhaltung.

5. Die qualitative Geheimhaltung

Die qualitative Geheimhaltung erfolgt durch ein Verfahren, mit dem ggf. auch rein qualitative Basisdateien anonymisiert werden können. Basisdateien mit nur qualitativen Merkmalen (oftmals Registerdateien wie Einwohner- oder Unternehmensregister) zeichnen sich durch eine sehr große Anzahl an statistischen Objekten aus. Deshalb ist die Performance bei diesem Verfahren ein nicht zu vernachlässigendes Problem.

Die Lösung der qualitativen Geheimhaltung erfolgt auf der Randsummentabelle für alle qualitativen Merkmale im feinsten betrachteten Schlüssel. Diese Tabelle wird um die Anzahl, der der einzelnen Merkmalskombination zugeordneten Sätze der Basisdatei, erweitert. Das hat den Vorteil, dass in den qualitativen Feldern identische Sätze sofort erkannt werden (Anzahl>1). Quantitative Felder werden in dieser Stufe noch vernachlässigt. Die Sätze der Randsummentabelle werden sortiert, so dass möglichst wenige Schlüsseländerungen zwischen benachbarten Zeilen anzutreffen sind.

Mathematisches Modell:

Der Ansatz geht von der Bestimmung einer „optimalen Teilstichprobe" aus, wobei die zu übernehmenden Sätze mit einer unkritischen Häufigkeit (>=3) vorhanden sein müssen.

Die einzelnen Objekte der Originaldatei seien gruppiert (völlig identische Sätze zusammengefasst und die entsprechende Häufigkeit vermerkt).

Dann beschreibt

X - Vektor der Originalhäufigkeiten x_i der statistischen Objekte derAusprägungskombination i ;i=1,2,…,n (n-Anzahl der unterschiedlichen Ausprägungskombinationen)

Y - Vektor der dominanten statistischen Objekte y_i mit i=1,2,…,n
y_i=1 – wenn die Ausprägungskombination in einem Tabellenfeld Dominanz bewirkt
y_i=0 – wenn die Ausprägungskombination in keinem Tabellenfeld Dominanz bewirkt (zur Ermittlung dieses Vektors werden die quantitativen Felder noch herangezogen.)

R - Vektor aller Häufigkeitsfelder (Randsummen der Anzahl der Objekte) über alle zu kontrollierenden Tabellen
r_j ;j=1,2,…,k (k-Anzahl der Häufigkeitsfelder in allen zu kontrollierenden Randsummentabellen)

A - Ausprägungsmatrix – Blockmatrix mit Einheitsvektoren in den einzelnen zu kontrollierenden Blöcken A_t ;t=1,2,…,T (T- Anzahl der zu kontrollierenden Randsummentabellen)
($a_{t_{ij}}$ = 1 - Wenn die Ausprägungskombination i die Merkmale so besitzt, das es in Tabellenfeld j dargestellt wird.
$a_{t_{ij}}$ = 0 - Wenn die Ausprägungskombination i die Merkmale so besitzt, das es in Tabellenfeld j nicht dargestellt wird.)

$$A = \begin{Bmatrix} A1 \\ A2 \\ \hline AT \end{Bmatrix}$$

mit

$$A_1 = \left\{ \begin{array}{c|c|c} a_{111}=1 & a_{11j}=0 & a_{11n}=0 \\ a_{121}=0 & a_{12j}=0 & a_{12n}=0 \\ \hline a_{1i1}=0 & a_{1ij}=1 & a_{1in}=0 \\ \hline a_{1(m-1)1}=0 & a_{1(m-1)j}=0 & a_{1(m-1)n}=0 \\ a_{1m1}=0 & a_{1mj}=0 & a_{1mn}=1 \end{array} \right\}$$

Die Matrix A unterteilt sich in untereinander liegende Blöcke von aus Einheitsvektoren bestehenden Matrizen. Die Anzahl der Blöcke ist identisch mit der Anzahl der zu kontrollierenden Tabellen (siehe obere Abbildung).

In dieser Beispieldarstellung sind die Ausprägungskombinationen nach der 1. Randsummentabelle sortiert. Das 1. Objekt gehört somit in das 1. Tabellenfeld der 1. Tabelle usw. (m-Anzahl der Tabellenfelder der Tabelle 1).

So lässt sich der Zusammenhang von den statistischen Objekten zu den Randsummen darstellen als:

$$AX=R$$

Bei einer anonymen Datei muss gelten $x_i \varepsilon \{0,3,4,5,\ldots\}$. (Alle vorhandenen Objekte haben eine Häufigkeit von mindestens 3).

Bei der bisherigen Formulierung würden nur existierende Merkmalskombinationen in der Lösung vorkommen, da alle x_i real existierende Objekte sind. Um zu verhindern, das aus jedem Satz der Basisdatei geschlossen werden kann, dass die Merkmalskombination auch existieren muss, können plausible aber nicht vorkommende Merkmalskombinationen außerhalb von SAFE generiert und das Modell entsprechend erweitert werden. Die Anfangshäufigkeit x_i für diese Kombinationen ist natürlich 0.

Für Basisdateien mit quantitativen Feldern wird zusätzlich die Bedingung eingeführt, dass dominante Objekte möglichst nicht entfernt werden sollen:

$$x_i > 0 \,|\, y_i > 0 \quad \forall i, i = 1,2,\ldots,n$$

Diese Bedingung ist erforderlich, um qualitative Veränderungen bei großen statistischen Objekten zu vermeiden. Falls unter dieser Bedingung allerdings keine Lösung gefunden wird, muss sie vernachlässigt werden.

Da das obige Gleichungssystem in der Regel mit dieser Bedingung keine Lösung hat, ist ein Fehlervektor F (f_j ;j=1,2,…,k - Fehler im Tabellenfeld j) einzufügen.

Die Menge aller möglichen anonymen Lösungen beschreibt sich dann als:

$$AX + F = R$$

$$\sum_{i=1}^{n} x_i = g$$

mit g als Gesamtanzahl der Statistischen Objekte

$$x_i \in \{0,3,4,5,\ldots\}$$

$$i = 1,2,\ldots,n$$

$$j = 1,2,\ldots,k$$

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 den späteren Datennutzer 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) ein relativer Fehler von -100 % bzw. +200 % unumgänglich ist. Diesen Fehler für die gesamte Datei zu akzeptieren, würde aber zu unbrauchbaren Ergebnissen führen. Der Maximalfehler in den Randsummen erwies sich als brauchbares Optimalitätskriterium. Aber auch die Summe aller Abweichungen oder die Summe aller quadratischen Abweichungen wären denkbar.

Es ist somit eine Lösung gesucht, die für den kleinsten Maximalfehler existiert.

$$Z = \min\left(\max_j (abs(f_j)) \right)$$

$$AX + F = R$$

$$\sum_{i=1}^{n} x_i = g$$

$$x_i \in \{0,3,4,5,\ldots\}$$

$$i = 1,2,\ldots,n$$

$$j = 1,2,\ldots,k$$

Die Aufgabe besteht darin, durch Veränderung der Häufigkeit des Auftretens der Merkmalskombinationen die Fallzahlprobleme zu beseitigen. Dabei können die Fallzahlprobleme entweder auf eine Häufigkeit von 3 oder größer bzw. auf die Häufigkeit 0 geändert werden. 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 Randsummenfehler) und k+n+1 Nebenbedingungen (Randsummengleichungen, Mengeneinschränkung für x_i und Gesamtanzahl der Objekte). Aufgrund der n Nebenbedingungen zu x_i ist die Aufgabe nichtlinear und ganzzahlig. (Auch unter dieser Bedingung können noch mehrere Lösungen 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 7 Merkmalen mit zusammen 16 verschiedenen Schlüsselstufen und 149 817 zu kontrollierenden Randsummenfeldern) ergibt sich ein Speicherbedarf von $(n+k+1)^2$ 4Byte =233GB. Dieser müsste für einen schnellen Zugriff als Hauptspeicher verfügbar sein. Deshalb wurde hier ein numerischer Algorithmus gewählt, der sich durch eine schrittweise Annäherung an die Lösung auszeichnet. Es wird ein Algorithmus festgelegt, der von einer bekannten schlechten Lösung zu einer besseren Lösung führt. Das wiederholte Anwenden der Regeln führt dann zum Auffinden der gesuchten Lösung.

Lösungsweg:

Bestimmung der Startlösung:

Da zu Beginn keine anonyme Startlösung bekannt ist, gibt es zwei Möglichkeiten.

1. Bestimmung einer anonymen Startlösung durch einen einfachen Algorithmus (z. B. jeden 3. Satz der Basisdatei mit der Häufigkeit 3 nutzen). Von dieser Lösung muss dann bei den einzelnen Schritten eine schrittweise Verbesserung der Randsummenqualität erfolgen.
2. „Aufweichen der Lösungsmenge" – Es werden noch nicht vollständig anonymisierte Lösungen in die Lösungsmenge mit aufgenommen, damit sind die Originalhäufigkeiten bereits eine Startlösung. Der Algorithmus muss dann jedoch die Anzahl der noch vorhandenen Geheimhaltungsfälle mit in die Zielkriterien aufnehmen.

Nach verschiedenen Tests wurde der zweite Weg verwendet, weil so der Vorteil besteht, dass der Randsummenfehler in der Startlösung minimal ist. (Die Originalhäufigkeiten haben natürlich einen Randsummenfehler von 0).

Zielfunktion / Entscheidungsregel:

Die Zielkriterien waren im folgenden (in Anwendungsreihenfolge)

1. Minimierung des maximalen Fehlers in den Randsummen
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 Randsummen

Zwischen dem 1. und dem 2. Ziel besteht allerdings ein Widerspruch. Ein Durchführen von Gruppierungen zur Verringerung der Geheimhaltungsfälle muss z. B. bei der Startlösung zu Randsummenfehlern führen. Auch danach ist es nicht immer auszuschließen, das neue Fehler auftreten. Deshalb wird das erste Zielkriterium durch die Vorgabe einer Fehlerschranke erreicht. Beginnend mit einem Startfehler (in der Regel +-2) wird die Realisierung der Teilziele 2 bis 4 angestrebt. Die Fehlerschranke von +-2 ergibt sich, weil für kleinere Schranken die Lösbarkeit unter bestimmten Konstellationen nicht gegeben ist. Stammen z. B. in einer Schlüsselgruppierung alle 3 Einheiten aus verschiedenen Einzelschlüsseln, so kann eine Lösung nur erreicht werden, indem man einen Schlüssel als Repräsentant auswählt und diesen Schlüssel den anderen zuweist. Damit erhält diese Ausprägung jedoch eine Abweichung von +2 Einheiten (siehe Tabelle).

WZ93	Betriebe original	Betriebe anonym	Abweichung
DB1	3	3	±0
DB113	1	0	-1
DB121	1	3	+2
DB123	1	0	-1

Wenn Erfahrungen aus der Lösung gleichartiger Beispiele vorliegen, können auch größere Schranken vorgegeben werden.

Beim schrittweisen Durchlaufen der Datei werden alle diejenigen potenziellen Veränderungen durchgeführt, die die Ziele 2 bis 4 verbessern. Um Kompensationsveränderungen bei benachbarten Sätzen zu berücksichtigen, werden sequentiell gleitende Gruppen von 3 ggf. 4 Sätzen aus der Datei gezogen. Für diese Gruppen werden alle möglichen Veränderungskombinationen +-1, +-2 und ggf. +-3 gebildet. Die Bedingung, dass dominante Sätze nicht verschwinden sollen, wird bei den Veränderungskombinationen berücksichtigt. Die Auswahl einer durchzuführenden Veränderung erfolgt dann nach folgendem Schema:

1. Die Menge aller Veränderungskombinationen wird reduziert um die Kombinationen, die nach ihrer Realisierung mehr Geheimhaltungsfälle erzeugen würden, als vorher vorhanden waren.
2. Für diese Veränderungskombinationen wird die Veränderung der Ziele 2 bis 4 und die Verletzung des vorgegebenen Maximalfehlers getestet. Es werden dann nur noch die Veränderungskombinationen ausgewählt, die die Maximalfehlerbeschränkung nicht verletzen.
3. Die mögliche Veränderung wird bestimmt, indem man die Kombinationen auswählt, die das Ziel 2 am meisten verbessern (die meisten Geheimhaltungsfälle beseitigen). Verbleiben mehrere Kombinationen in der Auswahl, so wird das nächste Teilziel (bis zum 4. Teilziel) für die Entscheidung herangezogen. Es werden ggf. auch die bezüglich höherwertiger Teilziele neutralen Kombinationen ausgewählt, wenn keine Kombination dieses höherwertige Teilziel verbessert.

Die ausgewählte Kombination verbessert somit mindestens 1 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 sei eine Datei mit den qualitativen Merkmalen Wirtschaftsklassifikation und Region zu anonymisieren, als zu kontrollierende Randsummentabellen seien nur folgende eindimensionale Randsummen zu testen: WZ93 als 2,3 und 5 Steller und die Region für Berlin nach Stadtteil (Berlin-West, Berlin-Ost) und Bezirk. Mehrdimensionale Tabellen seien vernachlässigt. Die Tabellen enthalten die Abweichungen in der Anzahl der Betriebe, die sich im Verlauf der bisherigen Anonymisierung ergeben haben.

WZ93 2-Steller	Abw.	WZ93 3-Steller	Abw.	WZ93 5-Steller	Abw.
…	…	…	…	…	…
CA	-1	DA1	+3	DA189	+1
CB	+2	DA2	-1	DA191	+1
DA	+2	DB1	±0	DA196	-1
DB	-1	DB2	-1	DA197	+1
DC	±0	DC1	±0	DA200	±0
…	…	…	…	…	…

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

Bezirke	Abw.
…	…
03	+1
04	-2
05	+2
06	-1
…	…

Die aktuelle Maximalfehlerschranke sei ±3. Folgende Satzgruppe wird aktuell untersucht:

Satz in der Satzgruppe	WZ93	Bezirk	Stadtteil Berlin-…	Anzahl Betriebe	Veränderungsvarianten der Zeile
…	…	…	…	…	
1	DA191	03	Ost	3	-3, -1, +1
2	DA197	05	West	1	-1, +2
3	DA200	06	West	1	-1, +2
…	…	…	…	…	

Damit ergeben sich folgende Veränderungsvarianten:

Variante	Veränderung Satz … 1	2	3	Zielkriterien / Entscheidungskriterien: Veränderung der Anzahl an Geheimhaltungsfällen	Bleibt Veränderung innerhalb der maximalen Fehlerschranke?	Veränderung des Randabstandes der Lösung[1]	Veränderung des mittleren Randsummenfehlers	
1	-3	-1	0	-1	ja	0	0	
2	-3	-1	-1	**-2**	ja	-17	5	
3	-3	-1	+2	**-2**	ja	**4**	-2	auszuwählende Kombination
4	-3	+2	0	-1	nein			
5	-3	+2	-1	-2	nein			
6	-3	+2	+2	-2	nein			
7	-3	0	0	0	ja	3	-1	
8	-3	0	-1	-1	ja	-10	4	
9	-3	0	+2	-1	ja	-1	1	
10	-1	-1	0	0	ja	16	-8	
11	-1	-1	-1	-1	ja	3	-3	
12	-1	-1	+2	-1	ja	12	-6	
13	-1	+2	0	0	nein			
14	-1	+2	-1	-1	nein			
15	-1	+2	+2	-1	nein			
16	-1	0	0	+1				
17	-1	0	-1	0	ja	2	-2	
18	-1	0	+2	0	ja	-1	-1	
19	+1	-1	0	-1	ja	-8	2	
20	+1	-1	-1	**-2**	ja	-17	5	
21	+1	-1	+2	-2	nein			
22	+1	+2	0	-1	nein			
23	+1	+2	-1	-2	nein			
24	+1	+2	+2	-2	nein			
25	+1	0	0	0	nein			
26	+1	0	-1	-1	nein			
27	+1	0	+2	-1	nein			
28	0	-1	0	-1	ja	9	-3	
29	0	-1	-1	**-2**	ja	-2	0	
30	0	-1	+2	**-2**	ja	1	-1	
31	0	+2	0	-1	nein			
32	0	+2	-1	-2	nein			
33	0	+2	+2	-2	nein			
34	0	0	-1	-1	ja	-7	3	
35	0	0	+2	-1	nein			

1 siehe unten Teilziel: Maximierung der Möglichkeiten für weitere Veränderungen

Nach Entfernung der unmöglichen Veränderungen (mehr Geheimhaltungsfälle oder Überschreitung des Maximalfehlers sind hier grün unterlegt), wird aus den verbleibenden Möglichkeiten nach folgender Auswahl ausgewählt (jeweils fett dargestellt):

1. Die meisten Geheimhaltungsfälle beseitigen die Kombinationen 2,3,20,29 und 30 (jeweils 2).
2. Den größte Verbesserung des Randabstandes (geringste Verschlechterung) erzeugt dann die Kombination 3. Wären hier immer noch mehrere Kombinationen gleichwertig, wie z. B. bei Kombination 2 und 20, entscheidet der Einfluß auf den mittleren Randsummenfehler (Spalte 4 möglichst klein).

Die Veränderungen der Kombination 3 werden durchgeführt. Bei diesem Beispiel ist erkennbar, dass ggf. auch nicht geheimzuhaltende Fälle mit verändert werden, wenn es für das Gesamtproblem nützlich ist. Die Veränderungen von nicht geheimzuhaltenden Fällen werden aber erst geprüft, wenn sonst keine Lösung möglich ist (siehe weiter unten „Auswahltechniken der zu testenden Satzgruppen").

Danach wird eine neue Satzgruppe 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.

Teilziel: Maximierung der Möglichkeiten für weitere Veränderungen:

Bei der Auswahl aus mehreren Veränderungsmöglichkeiten, die die gleiche Anzahl an Geheimhaltungsfällen beseitigen, werden die Möglichkeiten für weitere Veränderungen berücksichtigt. Innerhalb der Satzgruppe können sich nicht alle Veränderungen in den Randsummenfehlern gegenseitig kompensieren, da die Sätze nicht vollständig identisch sind. Für einen einzelnen Geheimhaltungsfall in einer Basisdatei ist die Beseitigung nur dann gefährdet, wenn die Fehler in den Randsummen auf beiden Seiten (positive und negative Abweichungen) sich zu dicht am Rand befinden. Für die Beseitigung eines Geheimhaltungsfalles durch Verringerung der Häufigkeit darf keine Randsumme einen negativen Fehler haben, der gleich dem zulässigen Maximalfehler $-f_{max}$ ist. Analog ist die Beseitigung eines Geheimhaltungsfalles durch Erhöhung der Häufigkeit von 1 auf 3 nur dann möglich, wenn die Randsummenabweichung nach oben weder f_{max} noch f_{max} -1 beträgt. Weiterhin verhindert eine Randsummenabweichung von $>=f_{max}$ -2 bzw. $<=2-f_{max}$ die Korrektur einer Geheimhaltungs-entscheidung (Wechsel der Häufigkeit von 0 auf 3 und umgekehrt). Deshalb werden diese kritischen Randsummenfehler berücksichtigt, wenn eine Auswahlmöglichkeit unter mehreren Kombinationen besteht, die die gleiche Anzahl an Geheimhaltungsfällen beseitigen.

Eine einfache Fehlerfunktion zum „mittleren" Randsummenfehler kann diesem Ziel nicht gerecht werden, vor allem dann nicht, wenn auch noch versucht wird, mit zwei getrennten Maximalfehlerschranken für ein- und mehrdimensionale Randsummen zu arbeiten. Als Regel wird folgende Straffunktion S zur Messung des Randabstandes verwendet:

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

$$s_i = \begin{cases} 9 & ;\forall abs(f_i) = f_{max} \\ 3 & ;\forall abs(f_i) = f_{max}-1 \\ 1 & ;\forall abs(f_i) = f_{max}-2 \\ 0 & ;\forall abs(f_i) < f_{max}-2 \end{cases}$$

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

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, bei der Straffunktion sich nur auf die störenden Randsummen des gerade betrachteten Geheimhaltungsfalls zu konzentrieren. Eine „Bevorzugung" eines aktuellen Geheimhaltungsfalles führt bei einem Wechsel dieses aktuellen Geheimhaltungsfalles automatisch dazu,

dass die obige Regel verletzt ist und somit Schleifen nicht mehr ausgeschlossen werden können. Im obigen Fall hat jede Lösung für alle 3 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 erste Teilfunktion (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 zweite Teilfunktion (Straffunktion für in Nähe des Maximalfehlers liegende Randsummen) ist ebenfalls nach unten beschränkt. Sie ist 0, wenn keine Randsumme im Bereich $f_{max} \geq abs(f_i) \geq f_{max} - 2$ liegt.
  Die dritte Teilfunktion (mittlerer Randsummenfehler) hat ihr Minimum, wenn alle Randsummen fehlerfrei sind. Da für die zweite und dritte Teilfunktion keine Einschränkungen bezüglich Zulässigkeit gestellt werden, sind alle Kombinationen bezüglich dieser Teilfunktionen zulässige Lösungen.

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 und reagieren. Das wird bei diesem Verfahren mit dem „Gashebel" zulässiger Maximalfehler geregelt. Wird bei einem Durchlauf nicht ein geforderter Anteil von Geheimhaltungsfällen beseitigt, so erfolgt eine Vergrößerung des zulässigen Maximalfehlers. Damit entstehen wieder größere Freiräume, wodurch beim erneuten Durchlauf wieder mehr 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 jedoch zu Lasten der Qualität des Ergebnisses (höherer Maximalfehler).

Auswahltechniken der zu testenden Satzgruppen:

Die Entscheidung zur Beseitigung der Geheimhaltungsfälle durch Verändern der Häufigkeit wird nur dann verhindert, wenn die Veränderung mit einer Verletzung des Maximalfehlers einhergehen würde. Dann wird die Entscheidung vertagt. Durch nachfolgende Veränderungen kann es durchaus sein, dass bei einem erneuten Abtesten die dann vorhandenen Randsummen die Beseitigung ermöglichen, da sich die konkreten Randsummenfehler mit jeder Veränderung ebenfalls verändern können und auf die Schaffung von entsprechenden Freiräumen Wert gelegt wurde.

Durch diese gezielten Auswahltechniken sind folgende Probleme zu lösen:

- Gruppierung zu Häufigkeiten größer 2,
- automatische Erkennung des erforderlichen Maximalfehlers,
- Kontrolle und ggf. Korrektur von „alten" Geheimhaltungsentscheidungen.

Zu Beginn eines Anonymisierungslaufs ist der erforderliche Maximalfehler in der Regel nicht bekannt. (Bei mehrfachen Läufen für den gleichen Datenbestand beispielsweise Monatsreihen könnten Erfahrungen vorliegen.) Es muss aber ein Maximalfehler vorgegeben werden. Problematisch sind dabei zu eng gestellte Maximalfehlerforderungen. Das würde dazu führen, dass zuerst die homogenen Satzgruppen anonymisiert werden und im nachhinein immer inhomogenere Sätze übrig bleiben (einzelne gleichartige Satzgruppen dazwischen wurden ja entfernt). Bei einem ständigen Durchsuchen der Gesamtdatei würden mit großem Testaufwand die homogen Satzgruppen anonymisiert, während die inhomogenen immer weiter vertagt werden. Um die erforderlichen Maximalfehler 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 dazugenommen. Das „fast vollständig" ist erforderlich, um die Möglichkeit zu besitzen, strukturelle Ausreißer mit den kombinatorischen Möglichkeiten der Gesamtdatei zu entscheiden. Ist die Anonymisierung für diesen Teil nicht möglich, wird bereits ein höherer Maximalfehler akzeptiert. Eine Erhöhung des Maximalfehlers schafft in jedem Fall neue Spielräume, um weitere Veränderungen an geheimzuhaltenden Sätzen vorzunehmen. Somit kann auch bei sehr großen Dateien schnell der notwendige Maximalfehler erkannt werden und das Verfahren schneller laufen. Die verbleibenden Reste in der Datei werden regelmäßig wieder mitgetestet, da sich durch jede Veränderung auch die Randsummenfehler ändern und somit ständig neue Möglichkeiten existieren.

Beim Durchlaufen der Datei werden standardmäßig nur die noch existierenden Geheimhaltungsfälle betrachtet. Es werden immer 3 benachbarte Geheimhaltungsfälle getestet. Tritt trotzdem eine Stagnation des Verfahrens ein, so werden nacheinander andere Auswahlverfahren durchgeführt, wobei nach jedem Lauf getestet wird, ob die Stagnation noch vorhanden ist. Die weiteren Auswahlverfahren sind:

1. Die noch vorhandenen Geheimhaltungsfälle werden zusammen mit ihren physisch benachbarten Sätzen getestet. Hier kann es beispielsweise vorkommen, dass es möglich ist, einen Geheimhaltungsfall zu einer bereits vorhandenen 3er-Gruppe mit hinzuzunehmen.
2. Es werden alle noch vorhandenen Sätze getestet. Da es auch vorkommen kann, dass in der Nähe eines Geheimhaltungsfalles (2 Sätze davor und dahinter) alle Sätze entfernt wurden, werden im zweiten Versuch alle nicht mehr vorhandenen Sätze vernachlässigt und aus der verbleibenden Menge die Auswahlgruppen 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.
3. Nur als Notlösung (am Ende der Anonymisierung) wird ein Gesamtdurchlauf der Datei durchgeführt. Dann kann auch das Entfernen eines Satzes noch mal revidiert werden, wenn es dem Gesamtziel (siehe Entscheidungsregeln) dient. Dieses Auswahlverfahren ist am rechenaufwendigsten und wird nur dann

angewendet, wenn bereits der Gesamtbestand bearbeitet wird und nur noch vereinzelt Geheimhaltungsfälle existieren.

Zeitverlauf:
Das Verfahren zeichnet sich durch einen hyperbolischen Zeitverlauf aus. Einer sehr schnellen Anonymisierung des 1. Teils folgt eine sehr langsame Anonymisierung des 2. Teils. Nach 50 % der Gesamtlaufzeit sind in der Regel nur noch weniger als 5 % der Geheimhaltungsfälle zu bearbeiten. Für diese wird durch die Nutzung von aufwendigeren Auswahlalgorithmen mehr Zeit benötigt. Es wäre auch möglich, den Zeitverlauf linearer zu gestalten (Anpassung des Stagnationsmaßes). Das geht aber mit einer Erhöhung des Maximalfehlers im Ergebnis einher (bei Tests ergab sich ein ca. 25 % größerer Maximalfehler). Optimal wäre eine automatische Umschaltung des Verfahrens, wenn eine maximale Rechenzeit vorgegeben wird (Weiterentwicklungsvariante).

Eindimensionale Randsummen:
Da die Auffassung vertreten wird, dass die Wertigkeit eindimensionaler Randsummen höher einzuschätzen ist, wurde die externe Forderung postuliert, dass diese Randsummen einen geringeren Fehler aufweisen müssen. Dabei wird als Zielvorstellung der theoretische Minimalfehler von +-2 unterstellt. Für eine Minimalfehlervorstellung von weniger als 2 lassen sich Gegenbeispiele zeigen, wo die Forderung nicht realisierbar ist.

Auch im obigen Ansatz wurden deshalb unterschiedliche Fehlerschranken ein- und mehrdimensional unterstellt. Da eine ständige Restriktion von 2 jedoch für das numerische Verfahren zu eng ist, wurde folgende Regel eingebaut:

Es können dem Programm 3 Parameter übergeben werden. Die Parameter haben folgende Bedeutung:
1. Startfehler eindimensionaler Randsummen
2. Startfehler mehrdimensionaler Randsummen
3. maximaler Abstand der beiden Randsummenschranken

Bei Stagnation und der Entscheidung, die Randssummenschranken zu vergrößern, wird dann solange nur die mehrdimensionale Randsummenschranke aufgezogen, bis der maximale Abstand erreicht ist, danach erfolgt ein gleichzeitiges Aufziehen beider Randsummenschranken. (Hier wären auch andere Regeln denkbar.)

Um dem Wunsch nach einem Maximalfehler von 2 eindimensional nachzukommen, wird im Anschluss an die Anonymisierung folgende Aufgabe gelöst:

$$Z = (S_K(X)) \rightarrow \min$$
$$AX + F = R$$
$$\sum_{i=1}^{n} x_i = G$$
$$x_i \in \{0,3,4,5,\ldots\}$$
$$i = 1,2,\ldots,n$$
$$j = 1,2,\ldots,k$$

Es wird eine neue Straffunktion S_K für die Lösungskorrektur eingeführt. Diese Funktion soll bewirken, dass unter den erreichten Maximalfehlerschranken die Lösung so verändert wird, dass die eindimensionalen Fehler größer 2 schrittweise abgebaut werden.

Zuerst werden die maximalen eindimensionalen und mehrdimensionalen Fehler bestimmt. Erhält man einen eindimensionalen Fehler größer als 2, so wird für die Korrektur ein um 1 verkleinerter maximaler eindimensionaler Fehler $f_{max,eindim}$ vorgegeben. Der maximale mehrdimensionale Fehler $f_{max,mehrdim}$ wird als Restriktion für die mehrdimensionalen Randsummen vorgegeben

Die Funktion sieht ungefähr folgendermaßen aus:

$$S_K = \sum_{i=1}^{k} s_{Ki}$$
$$s_{Ki} = \begin{cases} 100000 & ;\forall((abs(f_i) = f_{max,\,eindim} + 1) \wedge (i <= \text{eindim})) \\ 9 & ;\forall((abs(f_i) = f_{max,\,eindim}) \wedge (i <= \text{eindim})) \\ 3 & ;\forall((abs(f_i) = f_{max,\,eindim} - 1) \wedge (i <= \text{eindim})) \\ 1 & ;\forall((abs(f_i) = f_{max,\,eindim} - 2) \wedge (i <= \text{eindim})) \\ 0 & ;\forall((abs(f_i) \leq f_{max,\,eindim} - 2) \wedge (i <= \text{eindim})) \\ 9 & ;\forall((abs(f_i) = f_{max,\,mehrdim}) \wedge (i > \text{eindim})) \\ 3 & ;\forall((abs(f_i) = f_{max,\,mehrdim} - 1) \wedge (i > \text{eindim})) \\ 1 & ;\forall((abs(f_i) = f_{max,\,mehrdim} - 2) \wedge (i > \text{eindim})) \\ 0 & ;\forall((abs(f_i) \leq f_{max,\,mehrdim} - 2) \wedge (i > \text{eindim})) \end{cases}$$

Es wurde bei dieser Schreibweise unterstellt, dass alle eindimensionalen Randsummen vor die mehrdimensionalen sortiert sind und *eindim* die Anzahl der eindimensionalen Randsummentabellen darstellt.

Die Maximalstrafe von 100000 ergibt sich als hinreichend großer Wert für 10*(Anzahl der Randsummen). Es soll ein eindimensionales Problem auch dann beseitigt werden, wenn dabei alle anderen Randsummen auf den Rand verschoben werden. Alle anderen Fehler im Randbereich werden analog der Straffunktion im Anonymisierungsschritt bestraft. Der Wert der Maximalstrafe ist somit bei Bedarf anzupassen.

Mit dieser Straffunktion und dem mittleren Randsummenfehler als Zweitkriterium werden wieder schrittweise die möglichen Veränderungen (ohne Erzeugung von neuen Geheimhaltungsfällen) getestet. Das Ziel ist erreicht, wenn es eindimensional keine Randsumme mehr gibt, die den aktuellen eindimensionalen Maximalfehler besitzt. Dann wird der eindimensionale Maximalfehler um 1 verkleinert und die Aufgabe neu gelöst. Tritt vorher eine Stagnation ein, so könnte dann nur durch eine weitere Vergrößerung des mehrdimensionalen Maximalfehlers neuer Spielraum geschaffen werden. Hier entsteht ggf. ein Zielkonflikt. Beim Berliner Einwohnerregister (Bezirk 17) trat beispielsweise der Fall auf, dass eine Lösung aus dem Anonymisierungsschritt mit einem Fehler von maximal 3 eindimensional und 7 zwei- und dreidimensional erreicht wurde. Von den 69 Dreien unter den 8 253 eindimensionalen Randsummen konnten nicht alle beseitigt werden. Das war erst nach einem schrittweisen Aufziehen des mehrdimensionalen Randsummenfehlers auf letztendlich 10 möglich. Es mussten 112 mehrdimensionale Randsummen mit einem Fehler größer 7 hingenommen werden, um dem Anspruch an den eindimensionalen Fehler gerecht zu werden. Hier sollte der Fachbereich über den Bedarf einer solchen eindimensionalen Forderung wählen können. Tests haben unter anderem gezeigt, dass bei einer Korrektur bis zur Stagnation (ohne Aufziehen) nur noch vereinzelte eindimensionale Randsummen den gewünschten Fehler von maximal 2 überschritten (immer unter 5). Bei Dateien mit relativ wenigen qualitativen Merkmalen war die Herstellung eines qualitativen Fehlers von maximal 2 eindimensional in der Regel unproblematisch.

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 eigener Indextyp entwickelt. Dieser erinnert an einen grafischen Index, wie er für hierarchische Datenbanken verwendet wird. Er hat neben dem schnelleren Zugriff auch den Vorteil, das er weniger Speicherplatz erfordert und somit auch größere Probleme vollständig im Hauptspeicher realisierbar sind. Die Vorbereitungsprogramme zur Ermittlung der Randsummen und des Indexes sind noch in Clipper geschrieben. Daraus resultiert z. Zt. noch ein großer Teil des Rechenbedarfs für die Datenvorbereitung, weil das Programm noch eine 16-Bit-Anwendung ist. Hier sind noch Rechenzeitgewinne durch eine Umstellung auf eine schnellere Programmiersprache möglich.

6. Die quantitative Geheimhaltung

Mit der qualitativen Geheimhaltung hat man den Teil der qualitativen Merkmale der anonymen Basisdatei bestimmt. Um die quantitativen Merkmale zu bestimmen, wird als nächstes eine Zuordnung der Originalsätze vorgenommen.

Zuordnung der Lösung

Die Zuordnung soll natürlich so erfolgen, dass möglichst wenig qualitative Veränderungen gegenüber der Originaldatei erforderlich sind.

Die Basisdatei wurde sortiert und Sätze mit gleichen Ausprägungskombinationen wurden im vorigen Abschnitt zusammengefasst. Die daraus erhaltene Auswertungstabelle hat folgendes Aussehen: Alle qualitativen Merkmale werden in der feinsten Schlüsselstufe gespeichert. Die Häufigkeit der Ausprägungskombination wird als separates Feld (Vektor X^o) dahinter gespeichert. Im Ergebnis der im vorigen Abschnitt beschriebenen qualitativen Geheimhaltung wurde ein anonymer Häufigkeitsvektor (X^a) für die einzelnen Ausprägungskombinationen erhalten, der nur noch Häufigkeiten >=3 oder 0 enthält. Um diese Spalte der Zielhäufigkeit und um eine Arbeitsspalte für die aktuell zugeordnete Häufigkeit wird die Auswertungstabelle erweitert. Die Basisdatei wird um eine Hilfsspalte (Schalter zugeordnet; Startwert: nein) erweitert.

Unter den quantitativen Merkmalen wird ein dominierendes Merkmal für die Größenbestimmung der Unternehmen festgelegt (z. B. Beschäftigte). Dieses Merkmal dient zur Vergabe einer Priorität beim Wechsel von Zuordnungen der qualitativen Merkmale.

Beispiel: Es bestehen mehrere Unternehmen der gleichen Branche in verschiedenen Stadtbezirken. In zwei Ostberliner Bezirken existiert nur je ein Unternehmen. Im Zuge der qualitativen Geheimhaltung wurde festgelegt, dass ein Unternehmen einem anderen Ostberliner und das andere einem Westberliner Bezirk zugeordnet wird. Dann entscheidet die „Größe des Unternehmens", dass das kleinere Unternehmen den größeren „Sprung" macht und einem Westberliner Stadtbezirk zugeordnet wird.

Die Zuordnung der Originalsätze zu den erhaltenen Ausprägungskombinationen erfolgt deshalb nach folgender Regel:

1. Es wird eine Prioritätenreihenfolge der zu berücksichtigenden qualitativen Merkmale und ihrer Schlüsselstufen festgelegt. Aus dieser Prioritätenreihenfolge ergibt sich ein Gesamtschlüssel für die einzelnen Sätze der Basisdatei durch Aneinanderkettung der einzelnen Merkmalsschlüssel. Die Gesamtlänge dieses Schlüssels wird als Startlänge des Schlüssels festgelegt.
2. Die Basisdatei wird in der Reihenfolge des Gesamtschlüssels in der festgelegten Länge sortiert. Bei gleichen Schlüsselausprägungen erfolgt eine absteigende Sortierung nach dem dominierenden quantitativen Merkmal.
3. Die Basisdatei wird sequentiell durchlaufen. Für jeden noch nicht zugeordneten Satz der Basisdatei wird überprüft, ob es einen adäquaten Satz in der Ausprägungstabelle gibt, der im Bereich des aktuell kontrollierten Gesamtschlüssels identisch ist, und der in der Hilfsspalte der zugeordneten Sätze noch eine Menge enthält, die geringer ist als die zu erreichende Zielhäufigkeit. Ist das der Fall, werden alle qualitativen Eigenschaften der Ausprägungstabelle übernommen und der Satz in der Basisdatei als zugeordnet markiert. Die Menge der zugeordneten Sätze in der Arbeitstabelle wird um 1 erhöht.
4. Nach dem Durchlauf der Datei wird die Schlüssellänge des Gesamtschlüssels so verkürzt, dass eine weitere Merkmalstufe bzw. Merkmal aus dem Schlüssel entfällt. Sind noch nicht alle Sätze zugeordnet, wird mit Schritt 2 weitergearbeitet, ansonsten ist die Zuordnung beendet.

Die Sätze der Basisdatei werden durch diesen Algorithmus so den erhaltenen Ausprägungskombinationen zugeordnet, dass die einzelnen Objekte mit der größten Ähnlichkeit und bei gleich ähnlichen Sätzen zuerst die größten Objekte zugeordnet werden. Die größten qualitativen Veränderungen sollen also mit möglichst kleinen Unternehmen vorgenommen werden. Die eingangs festgelegte Option, dass dominante Sätze vorher bei der qualitativen Geheimhaltung markiert wurden, führte automatisch dazu, dass diese Sätze in ihrer originalen Struktur möglichst erhalten bleiben können. Nur bei wenigen Ausnahmen (widersprüchliche Dominanzen) ist das nicht der Fall.

Beispiel (widersprüchliche Dominanz): In einer Branche mit drei Unternehmen befinden sich alle Unternehmen in verschiedenen Regionen (Bezirken). Bei den Unternehmen dominiert eines beispielsweise den Umsatz und die Beschäftigten, während ein anderes den Auslandsumsatz dominiert, da das große Unternehmen kaum Export hatte. Die qualitative Anonymisierung muss dazu führen, dass die Unternehmen in einer Region zusammengefasst werden. Damit ist jedoch einer der beiden Dominanzfälle in seinen qualitativen Merkmalen nicht zu erhalten, wenn man andererseits den Fehler bei der Anzahl der Unternehmen nicht auf 3 erhöhen will.

Mit dieser Lösung werden jetzt wieder alle potenziellen Randsummentabellen berechnet. Dabei werden die originale und die anonyme Lösung parallel gehalten. Die bisherigen Schritte sichern bisher nur die Fallzahlprobleme und die Randsummenprobleme. Dominanzen sind dabei immer noch möglich. Die Auswertungstabellen werden deshalb für alle qualitativen Felder um die Werte originaler Wert, anonymer Wert, unzulässiger Bereich unten und unzulässiger Bereich oben erweitert. Die Werte des unzulässigen Bereiches sind natürlich nur dann belegt, wenn auf dem Tabellenwert ein Geheimhaltungsfall auftrat.

Beseitigung der Dominanzprobleme

Im nächsten Schritt wird versucht, die Lösung so zu verändern, dass auch die letzten Geheimhaltungsprobleme beseitigt sind. Die dadurch erhaltene Lösung wäre dann die Grundlage für eine erste faktisch anonymisierte Basisdatei. Es ist deshalb so vorzugehen, dass die obige Randsummentabelle im feinsten Schlüssel derart verändert wird, dass kein Tabellenwert mehr innerhalb des unzulässigen Bereiches liegt. Als sekundäres Ziel sollen natürlich Fehler in den Tabellen minimal gehalten werden.

Die obige Randsummentabelle wird sequentiell durchlaufen, und für jede Zeile und ihre Entsprechung in aggregierten Tabellen wird getestet, ob die anonymisierten Werte innerhalb des unzulässigen Bereiches liegen und sie diesen Bereich wegen Dominanz oder Fallzahlprobleme auslösten. Wenn das der Fall ist, werden die erforderlichen Veränderungen nach oben und unten ermittelt. Liegt die Zeile in mehreren Tabellen im unzulässigen Bereich, werden die Bereiche ggf. aufgezogen. Es wird dann folgende Entscheidungsregel verwendet:

Variante	Satz liegt in der feinsten Randsummentabelle im unzulässigen Bereich	Satz liegt in mindestens einer aggregierten Randsummentabelle im unzulässigen Bereich	Veränderung des quantitativen Wertes
1	wegen Dominanz oder Fallzahl	nein	an die obere Grenze des Bereiches der feinsten Tabelle
2	wegen Dominanz oder Fallzahl	wegen Dominanz durch anderen Satz	an die untere Grenze des Bereiches der feinsten Tabelle
3	wegen Dominanz oder Fallzahl	wegen Dominanz durch diesen Satz	an die obere Grenze des Bereiches der feinsten Tabelle
4	wegen Fallzahl	wegen Fallzahl	an die obere Grenze des Bereiches aller Tabellen
5	nein(weil der dominierende Satz qualitativ verändert wurde)	wegen Dominanz durch diesen Satz	an die untere Grenze des Bereiches aller Tabellen

Damit werden nur die Geheimhaltungsfälle verursachenden Sätze bearbeitet. Nach 2 bis 3 Durchläufen sind in der Regel alle Geheimhaltungsfälle beseitigt.

Optimierung der Lösung

Wird die Qualität der Lösung an der Reproduktion der Auswertungstabellen gemessen und nicht primär an der Ähnlichkeit zu den Originalsätzen, schließt sich hier eine Lösungsoptimierung an. Es wird für alle Randsummentabellen der Fehler in einer tabellenübergreifenden Fehlerfunktion abgebildet (z. B. die Summe der Quadrate aller Abweichungen zwischen Original und anonym, bei unzulässigen Bereichen bis zur Grenze des Bereiches). Im zweiten Schritt wird aus dieser Funktion der Einfluss eines einzelnen Zellwertes separiert und als Teilfunktion herausgetrennt. Durch Bestimmung der Ableitung der Teilfunktion und ein Nullsetzen der Ableitung erhält man eine Möglichkeit, die notwendige Änderungsrichtung fürs partielle Minimum zu bestimmen.

Dieser Algorithmus wird für jeden Wert der Tabelle der feinsten Randsummen automatisiert.

In den ersten Durchläufen werden diese Veränderungen nur dort angewandt, wo der Tabellenwert sowohl in der feinsten Tabelle als auch in allen Aggregaten in die gleiche Richtung vom Original abweicht. Ist so keine Verbesserung der Lösung mehr erreichbar, werden auch andere Werte verändert, bis bei einem Durchlauf „numerisch keine“ (kaum eine) Verbesserung mehr erreichbar ist. Man erhält so einen Satz von kompatiblen, anonymisierten Auswertungstabellen, die möglichst originalgetreu die Realität reproduzieren.

Die Notwendigkeit dieses Optimierungsschrittes hängt in starkem Maße vom Verwendungszweck ab. Bei dem bisher angestrebten Ziel, einen Datenkörper zu haben, mit dem konsistent anonymisierte Tabellen erzeugt werden können, ist er jedoch sehr hilfreich.

Das Erstellen der optimierten Basisdatei

In einem letzten Schritt muss die zuvor erhaltene und zugeordnete anonymisierte Basisdatei an diese feinste Randsummentabelle angepasst werden. Gleichzeitig ist die Identität unter den Sätzen so herzustellen, dass jeder Satz zu mindestens 2 weiteren identisch ist. Aus der ggf. erhaltenen Abweichung zwischen dem originalen und dem anonymisierten Wert in der feinsten Randsummentabelle ergeben sich ggf. erforderliche Veränderungen der Einzelsätze. Dabei ist zu unterscheiden, dass nur die Änderungen, die durch das „Nachjustieren“ der Lösung entstanden sind, verteilt werden müssen. Änderungen, die durch dass qualitative Verändern von statistischen Objekten entstanden (z. B. Zusammenlegen mehrerer Bezirke), bleiben unberücksichtigt. Diese erforderlichen Veränderungen lassen sich als Abstand zwischen der optimierten, anonymen Lösung und der nicht optimierten, anonymen Lösung (durch neues Aggregieren der anonymisierten Basisdatei) erkennen. Diese Veränderungen werden bisher relativ gleich auf die einzelnen Unternehmen aufgeteilt. Es wären hier aber auch andere Regeln denkbar, wie beispielsweise die Abweichung ausschließlich am größten Objekt (hatte den Geheimhaltungsaufwand ausgelöst) zu kompensieren. Danach werden die Unternehmen einer Gruppe nach ihrem dominierenden, quantitativen Merkmal absteigend sortiert und immer drei benachbarte, quantitative Werte durch ihren Durchschnitt ersetzt. Bei der letzten Gruppe kann es im Extremfall zur Bildung einer Gruppe aus fünf identischen Sätzen kommen. Da die Werte in der Regel ganzzahlig sein müssen, werden die Rundungsfehler an den Randsummen vermerkt und möglichst versucht, sie in der nächsten Gruppe wieder auszugleichen.

Im Ergebnis hat man eine neue anonymisierte Basisdatei erhalten, bei der immer mindestens drei Sätze identisch sind. Gleichzeitig sind alle potenziellen (berücksichtigten) Auswertungstabellen mit einem minimalen Fehler in der Firmenanzahl und möglichst originalgetreuen quantitativen Werten reproduzierbar. Die eingangs erläuterten möglichen Geheimhaltungsfälle sind ausgeschlossen, was zu teilweise größeren Abweichungen bei einzelnen quantitativen Werten führt. Diese sollten sich jedoch bei Bildung von höher aggregierten Tabellen kompensieren.

7. Zusammenfassung und Ausblick

Mit dem Verfahren SAFE steht eine Methode zur Verfügung, die es ermöglicht, die grundlegenden Geheimhaltungsansprüche sowohl für aus der Datei erstellte Tabellen als auch für Einzeldaten zu sichern. Gleichzeitig bleibt die flexible Auswertbarkeit der Einzeldaten gewährleistet.

Gerade in Bezug auf die Erstellung von Scientific-Use-Files, wie sie neuerdings verstärkt von der Wissenschaft gefordert werden, erweisen sich die Mikroaggregationsverfahren als brauchbar. Hier sind jedoch bei SAFE noch Anpassungen der Zielfunktionen/Teilschritte erforderlich, weil das Verfahren an der Qualität der erzeugbaren Auswertungstabellen optimiert wurde, was teilweise für die Qualitätsansprüche an den Einzeldaten kontraproduktiv ist.

Quellennachweis:

[1] Gesetz über die Statistik für Bundeszwecke (Bundesstatistikgesetz – BStatG) vom 22. Januar 1987 (BGBl. I S. 462, 565), zuletzt geändert durch Artikel 16 des Gesetzes vom 21. August 2002 (BGBl. I S. 3322).

[2] 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.