Wahrung der Geheimhaltung

sensibler Daten in mehrdimensionalen Tabellen

mit dem

Quaderverfahren

Rüdiger Dietz Repsilber

Stand: 23. 10. 2003

Landesamt für Datenverarbeitung und Statistik Nordrhein-Westfalen
Postfach 10 11 05, 40002 Düsseldorf
Mauerstraße 51, 40476 Düsseldorf
Telefon: 0211 9449-01
Telefax: 0211 442006
Internet: http://www.lds.nrw.de
E-Mail: poststelle@lds.nrw.de

© Landesamt für Datenverarbeitung und Statistik NRW,
Düsseldorf, 2003
Für nicht gewerbliche Zwecke sind Vervielfältigung und unentgeltliche Verbreitung, auch auszugsweise, mit Quellenangabe gestattet. Die Verbreitung, auch auszugsweise, über elektronische Systeme/Datenträger bedarf der vorherigen Zustimmung. Alle übrigen Rechte bleiben vorbehalten.

Zur Motivation

Die statistische Geheimhaltung, die Vermeidung der Offenlegung persönlicher Angaben in Veröffentlichungstabellen, ist eine fundamentale Aufgabe jeder Statistiken erhebenden und verbreitenden Institution, weil damit die für die Aussagefähigkeit der Daten unabdingbare Vertrauensbasis geschaffen wird. Andererseits ist mit dem Schutz persönlicher Daten gegen ihre Offenlegung untrennbar ein Informationsverlust verbunden, der die Aussagefähigkeit der veröffentlichten Statistik – wenn auch auf kontrollierbare Weise - einschränkt. Die Maxime muss daher sein, soviel Offenlegung wie möglich und nur soviel Geheimhaltung wie unbedingt nötig. So zu verfahren ist umso wichtiger, als diejenigen, die zu diesen Statistiken berichten, häufig auch zum Kreis der diese Statistiken Nachfragenden gehören, so dass ein wechselseitiges Interesse an einer möglichst optimalen Datensicherung besteht.

Als weiterer vertrauensbildender Maßnahme kommt der Offenlegung der angewendeten Geheimhaltungsverfahren selbst eine große Bedeutung zu. Diese Verfahren sollten daher ganz gezielt so entwickelt werden, dass sie sich – zumindest im Prinzip – allgemeinverständlich vermitteln lassen. Dabei steht nicht nur der Nutzer der hinsichtlich der statistischen Geheimhaltung gesicherten Tabellen im Blickfeld, sondern auch der die Statistiken vertreibende Fachstatistiker, weil er die nötigen Sicherungsmaßnahmen gegenüber dem Nutzer überzeugend vertreten können soll, was wiederum der Akzeptanz zugute kommt.

Gegenstand der statistischen Geheimhaltung sind nach mehreren Kriterien gegliederte oft mehrfach durch Zwischensummen unterteilte Statistiktabellen. In solchen fein gegliederten Tabellen treten viele Werte auf, die einzelnen Berichtenden zugeordnet werden können und die z.B. durch Sperren geheimgehalten werden müssen. Die Unterdrückung solcher Werten bezeichnet man als primäre Geheimhaltung. Das Sperren von sensiblen Werten allein genügt jedoch nicht, um sie vor zu genauer Berechnung mit Hilfe noch offener Tabellenwerte mittels der Tabellen-Summen-Beziehungen und einem gewissen, bei den Tabellennutzern zu unterstellenden Vorwissen über die Tabellenwerte zu schützen. Die Verhinderung der zu genauen Rückrechnung primär geheimer Werte aus noch offenen Tabellenwerten – ggf. unter Zuhilfenahme von beim Tabellennutzer zu unterstellendem Vorwissen - wird als sekundäre Geheimhaltung bezeichnet.

Die sekundäre Geheimhaltung hat ihre Ursache in der Veröffentlichung von Summen- und Zwischensummenwerten. Ohne Angabe solcher Summenwerte wäre alleine die primäre Geheimhaltung bereits hinreichend für den Schutz von persönlichen Angaben. Andererseits kann der Tabellennutzer im Nachhinein keine Summen oder Zwischensummen berechnen, wenn zu diesen gesperrte Werte beitragen. Erst die sekundäre Geheimhaltung, die Sperrungen auf den untersten Aggregationsniveaus setzt, macht die Veröffentlichung solcher Summenwerte überhaupt möglich.

Ein besonders einfaches Verfahren zur sekundären Geheimhaltung, das diese Anforderungen erfüllt, ist das Quaderverfahren, das auch Hauptthema dieser Dokumentation ist. Es dient hier zugleich als Vehikel, um damit die recht komplexen Geheimhaltungsstrukturen zu transportieren und verständlich zu machen.

Teil I: Grundlagen

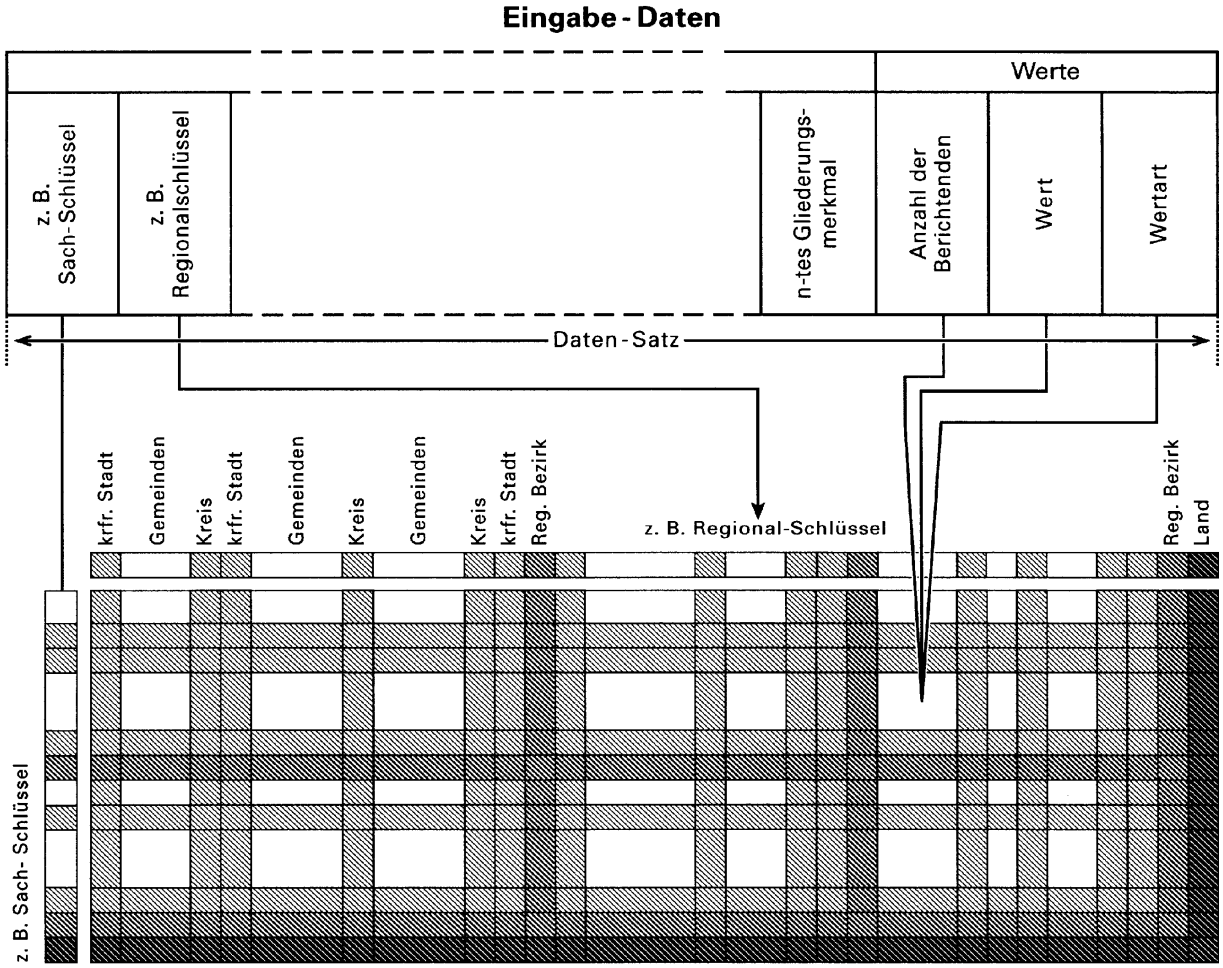
Einführung

Im Folgenden wird nach kurzer Beschreibung der Datengrundlage und der primären Geheimhaltung das seit einigen Jahren bei vielen Statistiken im Landesamt für Datenverarbeitung und Statistik Nordrhein-Westfalen und auch in anderen Bundesländern eingesetzte Quaderverfahren zur Wahrung der Geheimhaltung in aggregierten Daten dargestellt. Das Verfahren sichert nach mehreren Merkmalen gegliederte, auch mehrfach durch Zwischensummen unterteilte Tabellen gegen zu genaue Rückrechnung ihrer primär gesperrten Werte durch zusätzliche Sperrungen (Sekundärsperrungen) von Tabellenfeldern. Es bietet Intervallschutz für die primär gesperrten Werte, d.h. es verhindert, dass ein primär gesperrter Wert genauer schätzbar ist, als es ein vom Anwender vorgegebenes Intervall um den geheimen Wert erlaubt. Das Quaderverfahren wurde insbesondere zur Behandlung sehr umfangreicher Tabellen (z.B. 1 000 000 Tabellenfelder) konzipiert.

Als Eingabedaten benötigt das Quaderverfahren Tabellendaten. Man spricht von n-dimensionalen Tabellen, wenn diese nach n verschiedenen Merkmalen (z.B. Sach- und Regionalschlüssel) gegliedert sind. Jedes Tabellenfeld einer n-dimensionalen Tabelle kann durch einen Datensatz beschrieben werden, der die n Gliederungsmerkmale, die Anzahl der Berichtenden, den Wert des Veröffentlichungsmerkmals als Summe der Einzelmerkmalswerte und den Wertartschlüssel, d.h. die Kennzeichnung, ob der Wert geheim zu halten ist oder nicht, umfasst. Diese Datenbeschreibung deckt die Mindestanforderungen an das Quaderverfahren ab. Dazu gehört bereits die Berücksichtigung des Vorwissens, dass es sich um eine sogenannte positive Tabelle handelt, dass also keine negativen Tabellenwerte auftreten. Muss darüber hinaus unterstellt werden, dass die Tabellenwerte vom Nutzer sogar bis auf Schätzintervalle eingegrenzt werden können, so müssen jedem Datensatz noch ein unterer und ein oberer Schätzfehler hinzugefügt werden. Außerdem kann man diesen Datensatz noch durch ein weiteres Wertefeld zur externen Gewichtung ergänzen, um damit die Auswahl von Sekundärsperrungen zu steuern. Wenn nicht ausdrücklich erwähnt, wird im Folgenden von der Mindestanforderung ausgegangen.

Die Abbildung „Eingabedaten“ zeigt in ihrem oberen Teil den Datensatz einer n-dimensionalen Tabelle. Im unteren Teil der Abbildung ist eine Tabelle für den Fall zweier Gliederungsmerkmale (z.B. ein Sach- und ein Regionalschlüssel) als zweidimensionales Zahlentableau in Gestalt eines Rechtecks dargestellt, mit einer Unterteilung durch schraffierte Zeilen oder Spalten, die die Zwischen- bzw. Randsummenzeilen oder -spalten symbolisieren. Die unterschiedlich dunkle Schraffur gibt das unterschiedliche Aggregationsniveau an. So werden beispielsweise die durch den Sachschlüssel indizierten Zeilen des untersten Aggregationsniveaus, etwa der 4-Steller mit hellster (weil ohne) Schraffur zu ihrem jeweils darunter liegenden dunkler schraffierten 3-Steller aufsummiert; die Dreisteller ihrerseits zu ihren Zweistellern (mit zweitdunkelster Schraffur), die schließlich noch zum Einsteller mit dunkelster Schraffur entsprechend der höchsten Verdichtung zusammengefasst werden. Die Aggregierung der im Schaubild „Eingabedaten“ durch eine regionale Gliederung ausgeführten Spaltengliederung erfolgt von links nach rechts, indem die Werte der Gemeinden zu ihren Kreiswerten, die Kreiswerte und die Werte der kreisfreien Städte zu

ihren Regierungsbezirkswerten und die Regierungsbezirkswerte schließlich zum Landeswert aufaddiert werden. Die als Vorspalte bzw. Kopfzeile angefügten Indexleisten nehmen die Sach- bzw. Regionalschlüssel auf.



Die hier vorgestellte Tabellendefinition schließt sogenannte Kontingenztabellen mit ein, wenn das in solchen Tabellen fehlende Merkmal Wert durch die Anzahl der Berichtenden ersetzt wird.

Dieser Datenbestand ist - wie im Schaubild „Eingabedaten“ angedeutet - bezüglich jedes Gliederungskriteriums so sortiert, dass die höheren Aggregate den niedrigeren, aus denen sie bestehen, nachfolgen. Die Summen- und Zwischensummenstruktur legt eine Schlossdatei fest, die durch Zuordnung der Gliederungsmerkmale zu ihren Aggregationsniveaus in Verbindung mit der Sortiervorschrift Auskunft darüber gibt, wie Tabellenwerte und Berichtendenzahlen zu Zwischen- bzw. Randsummen aufaddiert wurden. In NRW ist beispielsweise das „Schloss“ der regionalen Gliederung gegeben durch die Zuordnungstabelle: alle Gemeindeschlüssel zur ersten Aggregationsstufe, alle Kreisschlüssel und Schlüssel der kreisfreien Städte zur zweiten Aggregationsstufe, alle Regierungsbezirksschlüssel zur dritten und der Landesschlüssel zur 4. Aggregationsstufe der regionalen Gliederung.

Mehrfach durch Zwischensummen unterteilte Tabellen können mit dem Quaderverfahren nur durch Überführung in zwischensummenfreie Tabellen mit genau einer Randsumme für jedes Gliederungskriterium bearbeitet werden. Das kann durch Aufteilung in eine Gesamtheit von Untertabellen geschehen, von denen jede eine zwischensummenfreie

Teilgesamtheit der Gesamttabelle mit nur einer Randsumme für jedes Gliederungskriterium umfasst. Solche Untertabellen können in Bezug auf die Wahrung der Geheimhaltung nicht unabhängig voneinander bearbeitet, sondern müssen aneinander abgeglichen werden, damit die in mehreren Untertabellen gemeinsam auftretenden Aggregate auch den selben Geheimhaltungsstatus haben. Dieses bisher immer noch praktizierte Vorgehen bietet nur einen notwendigen, keinen hinreichenden Schutz gegen zu genaue Rückrechnung der gesperrten Werte.

Um mit dem Quaderverfahren einen hinreichenden Intervallschutz zu erzielen, muss die gegebene Tabelle dazu durch Aufstocken der Dimension erweitert werden und zwar so, dass in der erweiterten Tabelle keine Zwischensummen mehr auftreten. Solche Tabellen werden hier als vollständige Tabellen bezeichnet und im Text näher beschrieben. Der ursprüngliche Dimensionsbegriff, der allein durch die Anzahl der Gliederungskriterien bestimmt ist, wird durch die Dimensionsaufstockung direkt mit der Aggregation der Tabellenwerte verknüpft: Jede in der zu sichernden Tabelle ausgewiesene Addition von Tabellenwerten zu einer Zwischen- oder Randsumme entspricht genau eine Gliederung bzw. Dimension der aufgestockten Tabelle, die zur Unterscheidung von der ursprünglichen Gliederung wegen ihrer Zwischensummenfreiheit als elementar bezeichnet werden soll. Bei mehrfach durch Zwischensummen unterteilten Tabellen kann erst die durch Dimensionsaufstockung umstrukturierte Tabelle mit dem Quaderverfahren hinreichend geschützt werden.

Viele Ansätze zur Wahrung der Geheimhaltung bei möglichst geringem Informationsverlust sind - obwohl mathematisch als Optimierungsproblem exakt lösbar[1)] - in Gestalt von Heuristiken realisiert, weil insbesondere bei umfangreichen Datenbeständen erhebliche Rechenzeiten eine exakte Lösung verhindern (siehe z.B. J. Geurts, Netherlands 1992). Im Falle von nicht negativen Tabellenwerten hat zwar L.H.COX, U.S. BUREAU of the CENSUS, Washington, 1992 beim internationalen Seminar zur statistischen Geheimhaltung in Dublin ein sogenanntes Netzwerkoptimierungsverfahren vorgeschlagen, das eine deutliche Reduktion der Rechenzeit für das lineare Optimierungsproblem bewirkt; dennoch muss das Geheimhaltungsproblem umfangreicher Tabellen in Unterprobleme unterteilt werden, weil die Rechenzeiten weit schneller als linear mit der Anzahl der Tabellenfelder zunehmen (siehe L.V. ZAYATZ, U.S. BUREAU of the CENSUS, 1992 Dublin und insbesondere J. Geurts, Netherlands 1992). Solche Verfahren sind allenfalls suboptimale Heuristiken, die in der Regel nicht einmal vollständig sicher sein müssen (siehe dazu auch das 6. Kapitel).

Außer den immensen Rechenzeiten zwingen aber noch andere praktische Gründe zur Übernahme von Heuristiken in die exakte lineare Optimierung, nämlich die Auswahl einer geeigneten Ziel- oder Kostenfunktion für die Sekundärsperrungen (siehe dazu insbesondere D.A. Robertson, Statistics Canada, "Automated Disclosure Control at Statistics Canada", vorgestellt auf dem zweiten internationalen Seminar über statistische Geheimhaltung in Luxemburg 1994). Wählt man beispielsweise die Summe der zusätzlich zum Schutze primär geheimer Werte zu unterdrückenden Werte als zu minimierende Funktion, erhält man oft eine unerwünscht hohe Zahl an Sekundärsperrungen (vergleiche dazu auch Abb. 1.2, Abschnitt 1.1.2). Umgekehrt werden, wenn die Anzahl der Sekundärsperrungen mini-

---

1) Als zu minimierende Zielfunktion wird dabei $\sum f(X) I(X)$ angenommen mit der Indikatorfunktion I (X) und der bei Unterdrückung von X dem Tabellennutzer vorenthaltenen Information f (X), I (X) = 1 für gesperrtes X und = 0 sonst**. Die sogenannte exakte Optimierung spart bisher eine exakte Behandlung von Einzelangaben aus!**

miert werden soll, besonders große Werte als Sperrpartner herangezogen. In o.g. Papier wird daher vorgeschlagen, zuerst mit einer degressiv wachsenden Kostenfunktion der Tabellenwerte X wie f(X) = lnX eine Vorauswahl der Sperrkandidaten zu treffen und dann in einem zweiten Schritt mit einer mit zunehmenden Werten X langsam abnehmenden(!) Funktion f(X) = lnX / X das Sperrmuster festzulegen.

Beim Quaderverfahren wird in erster Linie die Anzahl von Sekundärsperrungen minimiert und erst in zweiter Linie eine möglichst kleine Wertesumme zusätzlicher Sperrungen angestrebt. Dazu werden beide Kriterien praktisch einzeln abgefragt, so dass keine beide Kriterien gemeinsam berücksichtigende Zielfunktion wie z.B. $\sum f(X) * I(X) = \sum lnX * I(X)$ eingeführt werden muss, um hier die gewünschten Prioritäten einzuhalten. Zwar benutzen auch die derzeit das Quaderverfahren realisierenden EDV-Programme GHMITER und QUIT ebenfalls den Logarithmus der Werte an Stelle der Tabellenwerte selbst. Dies dient aber ausschließlich einer auch kleine Tabellenwerte noch genügend stark differenzierenden Werteklassierung, mit der die Information über die Wertattribute, offen, primär geheim, Einzelangabe, sekundär geheim etc. auf die Werte selbst übertragen wird. Mit dieser Werteklassierung wird Zugriffs- und dadurch Rechenzeit eingespart, ihre Rundungseffekte haben aber praktisch keinen Einfluss auf die Auswahl von Sperrpositionen. Die Zusammenlegung der Tabellenwerte und ihrer Attribute in Werteklassenstaffeln wird in einer Übersicht am Ende von 3.2.3 angedeutet und im fünften Abschnitt dann eingehender besprochen.

Trotz genau festgelegter Sperrprioritäten bietet das Quaderverfahren eine für praktische Anwendungen nützliche Steuerungsmöglichkeit, indem die Tabellenwerte vor der Bearbeitung mit dem Verfahren durch Gewichtung temporär geeignet modifiziert werden können, so dass sie auf Grund ihrer neuen Größe bevorzugt zu Sekundärsperrungen herangezogen oder besonders gemieden werden. Das bedeutet, dass der für die vorzunehmende Bewertung wesentliche Informations- oder Kostenbegriff, der bei Tabellen mit nicht negativen Werten im Allgemeinen direkt mit dem Tabellenwert selbst oder mit dessen Logarithmus in Zusammenhag gebracht wird, noch nachträglich, d.h. nach Erstellung der Tabelle nur für den Sicherungsvorgang uminterpretiert werden kann. Auf die Justierungsmöglichkeiten der Verteilung der Sekundärsperrungen durch Gewichtung von Werten wird im fünften Abschnitt und im Anhang bei der Behandlung von Beispielen noch gesondert eingegangen.

Wenn nicht gesondert angemerkt, behandelt diese Dokumentation ausschließlich sogenannte positive Werte-Tabellen. Damit sind Tabellen mit der oben angegebenen Struktur gemeint, die keine negativen Werte enthalten (siehe dazu die Abbildung „Eingabe-Daten“). Außerdem wird im Folgenden vorausgesetzt, dass zu voneinander verschiedenen Tabellenfeldern auch voneinander verschiedene Meldeeinheiten (Berichtende) gehören. Mit diesen Voraussetzungen werden bereits die weitaus meisten Tabellen in der amtlichen Statistik überdeckt, auf die Behandlung von Ausnahmen wird unter den Punkten 2.1.2.2 und 5.1.2.4 hingewiesen.

0. Anmerkungen zur primären Geheimhaltung

0.0 Allgemeines

Bei der Sicherung sensibler Tabellendaten gegen Offenlegung zu schützender Einzelangaben hat man in der amtlichen Statistik große komplex gegliederte Datenmengen zu bearbeiten (z.B. 1 000 000 Tabellenfelder), womit ein erheblicher personeller und organisatorischer Aufwand und darüber hinaus – wie bereits in der Einführung bemerkt – auch ein hoher Rechenzeitaufwand verbunden ist. Hilfreich war da zunächst eine Aufteilung des Geheimhaltungsproblems in die Geheimhaltung einzelner Tabellenwerte ohne Bezug auf ihre Tabellenumgebung, die sogenannte primäre Geheimhaltung, und in die Sicherung geheimzuhaltender Tabellenwerte gegen zu genaue Rückrechnung mit Hilfe von Tabellensummenbeziehungen, die sekundäre Geheimhaltung. Diese Aufteilung findet seither Anwendung bei den in der amtlichen Statistik weitverbreiteten Sperrverfahren und den daraus abzuleitenden, die Zwischen- und Randsummen unversehrt lassenden Verfälschungsverfahren (vgl. Abschnitt 4.1).

Mit der Aufteilung in primäre und sekundäre Geheimhaltung ging anfangs (zu Beginn der 80-er Jahre) auch eine Trennung der statistikspezifischen Geheimhaltungseigenschaften einher: In der primären Geheimhaltung waren diese Eigenschaften in Gestalt von Dominanzmaßen konzentriert, während der mehr formale, die Tabellenstruktur erfassende Teil der universeller anzuwendenden sekundären Geheimhaltung zukam. Die Grenzen zwischen diesen beiden Aspekten sind heute eher fließend, seit bekannt ist, wie der Intervallschutz in der sekundären Geheimhaltung die Genauigkeit der Schätzung eines Einzelwertes in seinem Tabellenwert maßgeblich mitbestimmt und primäre und sekundäre Geheimhaltung in der Sicherung von Mikrodaten mit Sekundärsperrverfahren vereinigt werden können (vgl. Abschnitt 4.2).

Die primäre Geheimhaltung schützt Einzelangaben unter Berücksichtigung von Vorwissen gegen zu genaue Rückrechnung aus dem Tabellenwert, zu dem sie beitragen (zur primären Geheimhaltung siehe auch S. Gießing, „Statistische Geheimhaltung“, Forum der Bundesstatistik, Bd. 31/1999). **Als Vorwissen wird das Wissen über die Tabelle ohne deren Kenntnis bezeichnet, das man beim Tabellennutzer unterstellen kann. Die sekundäre Geheimhaltung verhindert die zu genaue Rückrechnung der primär geheimen Werte aus noch offenen Tabellenwerten mit Hilfe der Tabellensummenbeziehungen und dem beim Tabellennutzer zu unterstellenden Vorwissen.**

Welche Tabellenfelder primär geheim zu halten sind, wird mit Hilfe von Geheimhaltungsregeln festgelegt. Dabei lassen sich zwei Gruppen von Geheimhaltungsregeln unterscheiden, je nachdem, ob Vorwissen über fremde, d.h. über nicht eigene Angaben des jeweiligen Berichtenden zu unterstellen ist oder nicht. Diese Einteilung ist beinahe deckungsgleich mit der Einteilung der primären Geheimhaltungsregeln nach eindeutiger (exakter) und näherungsweiser Rückrechenbarkeit.

0.1 Vermeidung eindeutiger Rückrechenbarkeit

Wird keinerlei Vorwissen unterstellt, das über eigene Angaben eines zum jeweiligen Tabellenwert Berichtenden hinausgeht, so sind die sogenannten Fallzahlregeln anzuwenden. Mit diesen Regeln ist nur eine eindeutige (exakte) Rückrechenbarkeit zu unterbinden. Im Prinzip verhindern Fallzahlregeln also nicht, dass eine zu sichernde Einzelangabe u.U. beliebig genau berechnet werden kann. Eine beliebig hohe Genauigkeit der Rückrechenbarkeit ist allerdings bei fehlendem Vorwissen über fremde Angaben völlig bedeutungslos, weil ein an Einzelwerten Interessierter zur eigenen Bestätigung seiner Rückrechnungsgenauigkeit zumindest rudimentäre Kenntnisse haben muss über Tabellenwerte und Anteile, mit welchen z.B. große Einzelangaben zum Tabellenwert beitragen und dieses Wissen ist hier voraussetzungsgemäß nicht vorhanden. **Bei fehlendem Vorwissen ist die Vermeidung der eindeutigen Rückrechenbarkeit für den Schutz von Einzelangaben gegen zu genaue Rückrechnung bereits hinreichend** (vgl. Abschnitt 4.2).

Bei den sogenannten Fallzahlregeln wird ein Tabellenwert geheimgehalten, wenn eine bestimmte Mindestfallzahl der zu diesem Tabellenwert beitragenden Berichtenden unterschritten wird. Bezeichnet N die Anzahl der Berichtenden, die zu dem betrachteten Tabellenwert beitragen, so lassen sich folgende Fälle unterscheiden:

N = 1: Die Geheimhaltung ist nicht gesichert, da nur ein Berichtender zum Tabellenwert beiträgt. Der Tabellenwert ist nach der Fallzahlregel für Tabellenwerte mit nur einem Berichtenden (Einzelangabe als Tabellenwert) primär geheim zu halten.

N = 2: Die Geheimhaltung ist nicht gesichert, da jeder der beiden zum Tabellenwert beitragenden Berichtenden die Einzelangabe des jeweils anderen als Differenzbetrag aus dem Tabellenwert und seiner eigenen Angabe exakt errechnen kann. Der Tabellenwert ist nach der Fallzahlregel für Tabellenwerte mit zwei Berichtenden primär geheim zu halten.

N > 2: Allgemein gilt: Wenn anzunehmen ist, dass ein „Datenangreifer“ N – 1 zusammengefasste Einzelbeiträge eines Tabellenwertes mit N Angaben kennt, so ist dieser Tabellenwert nicht gesichert, weil sich eine Einzelangabe durch Abzug der anderen vom Tabellenwert exakt berechnen lässt . Der Tabellenwert ist nach der Fallzahlregel für N Berichtende primär geheim zu halten.

In der deutschen amtlichen Statistik geht man im Allgemeinen davon aus, dass ab einer Fallzahl von N = 3 die Geheimhaltung hinsichtlich der Vermeidung einer exakten Offenlegung gesichert ist, weil der an der Offenlegung Interessierte nur seinen Einzelbeitrag kennt. Ab N = 3 Berichtenden wird also ein Tabellenwert aufgrund der Fallzahl nicht mehr primär gesperrt. Diese Regel ist als „Dreier-Regel“ bekannt.

Schon Fallzahlregeln, die bei mehr als zwei Berichtenden Primärsperrungen erfordern, setzen bereits ein Vorwissen über fremde Angaben voraus, müssen aber dennoch als Schutz gegen eine nur exakte Offenlegung interpretiert werden, weil sie, wie oben ausgeführt, primär nur eine exakte Offenlegung verhindern.

0.2 Vermeidung näherungsweiser Rückrechenbarkeit

Wenn außer der eindeutigen Offenlegung einer Einzelangabe, ganz allgemein auch die Offenlegung eines zu genauen Näherungswertes dieser Angabe verhindert werden soll, reichen die Fallzahlregeln für die primäre Geheimhaltung nicht aus. Die Möglichkeit zur näherungsweisen Offenlegung besteht z.B., wenn bekannt ist, dass der N Angaben umfassende Tabellenwert durch die n größten Einzelwerte dominiert wird, wobei n < N ist. Für die primäre Geheimhaltung können hier die sogenannten (n,k)-Dominanzregeln angewendet werden.

0.2.1 Die Dominanzregeln

0.2.1.1 (1,k)-Dominanzregel

Die (1,k)-Dominanzregel besagt, dass ein Tabellenwert primär geheim zu halten ist, wenn der Anteil des größten Einzelbeitrags X_1 an diesem Tabellenwert größer als der Parameter k ist, d.h. wenn gilt

$$X_1 > k\,X \qquad (0.1a)$$

Diese Regel stellt also sicher, dass bei veröffentlichten Tabellenwerten der Wert des größten Einzelbeitrags X_1 höchstens k% des zu veröffentlichenden Tabellenwertes X ausmacht.

Wenn man den Wert des größten Einzelbeitrags X_1 mit Hilfe des offenen Tabellenwertes X als $\hat{x}_1 = X$ schätzt, wird man ihn wegen (0.1a) gemäß $X_1 \leq k\,\hat{x}_1 \;\leftrightarrow\; (\hat{x}_1 - X_1)/X_1 \geq 1/k - 1$ um mindestens

$$M(1, k) = (1 - k) / k \;\leq\; (\hat{x}_1 - X_1) / X_1 \qquad (0.1b)$$

überschätzen. In den Statistischen Ämtern des Bundes und der Länder wird bisher häufig die (1,85)-Dominanzregel angewandt. Das ergibt eine „Mindest-Überschätzung“ des größten Einzelwertes M(1, 85) von $M(1, 85) = (1 - 0{,}85) / 0{,}85 = 17{,}6\,\%$.

Der Ansatz $\hat{x}_1 = X$ macht für den Tabellennutzer nur Sinn, wenn ihm der Parameter k unbekannt ist. Anderenfalls würde er bei offenem Tabellenwert X den Einzelwert $\hat{x}_1$ durch k*X sehr viel genauer schätzen können. D.h. Parameter zur Festlegung von Schutzintervallen sollten nicht veröffentlicht werden! Das gilt für die primäre wie für die sekundäre Geheimhaltung gleichermaßen: Die Kenntnis solcher Parameter bedeutet eine zusätzliche Vorinformation, die das Problem der Geheimhaltung – wie jede andere Vorinformation auch - wesentlich verschärft (vgl. auch 3.2.4).

Im nachfolgenden Beispiel werden die Wertangaben in Einheiten E angegeben, wo E z.B. für 1000 € stehen kann.

Beispiel : **Z**u beurteilen sei

der Tabellenwert $X = 100$ E

mit dem größte Einzelbeitrag $X_1 = 85$ E

und dem zweitgrößte Einzelbeitrag $X_2 = 10$ E

Wegen der Gültigkeit auch des Gleichheitszeichens in $85 \text{ E} / 100 \text{ E} \geq k = 0{,}85$ ist die Bedingung für eine Primärsperrung nach der (1, 85)-Dominanzregel gemäß (0.1a) gerade nicht erfüllt; der Tabellenwert braucht nicht primär geheimgehalten werden. Die hier vorliegende „Mindest-Überschätzung" des größten Einzelwertes beträgt $M(1, 85) = 17{,}6\ \%$ – wie oben für den gebräuchlichen Wert $k = 0{,}85$ berechnet.

0.2.1.2 (2,k)-Dominanzregel

Wesentlich genauer als nur mit dem Tabellenwert X kann der Berichtende des zweitgrößten Einzelbeitrages den größten Einzelwert X_1 schätzen, indem er seinen eigenen Beitrag X_2 vom Aggregatwert X abzieht: $\hat{x}_1 = X - X_2$. Wenn X_2 in obigem Beispiel 10 E beträgt, schätzt der Berichtende des zweigrößten Einzelwertes den größten Einzelwert als $\hat{x}_1 = 100 \text{ E} - 10 \text{ E} = 90 \text{ E}$ mit einem Fehler von $(90 \text{ E} - 85 \text{ E}) / 85 \text{ E} = 5{,}9\ \%$, unterschreitet also die durch die (1, 85)-Regel garantierte „Mindest-Überschätzung" von 17,6 % ganz beträchtlich. Die Situation lässt sich mit der Fallzahlregel für zwei Berichtende vergleichen, wobei hier ganz analog die Sicherung hinsichtlich der Dominanz mit der (1, k)-Regel mitunter auch keinen hinreichenden Schutz gegen zu genaues Schätzen des größten Einzelwertes gewährleistet, man braucht eine Dominanzregel für zwei dominierende Werte.

Die (2,k)-Dominanzregel berücksichtigt die beiden größten Einzelbeiträge X_1, X_2 eines Tabellenwertes X. Danach ist ein Tabellenwert X primär geheim zu halten , wenn der Anteil der Summe der zwei größten Einzelbeiträge, $X_1 + X_2$, am Tabellenwert X größer als der vorgegebene Parameter k ist:

$$X_1 + X_2 > k\,X \qquad (0.2a)$$

Nach der (2,k)-Dominanzregel können also nur noch Tabellenwerte veröffentlicht werden, deren Summenanteil der beiden größten Einzelwerte höchstens k% des Tabellenwertes X beträgt.

Der Melder des zweitgrößten Einzelwertes wird demnach den größten Einzelwert X_1 vermöge $\hat{x}_1 = X - X_2$ mindestens um

$$M(2, k) = (1 + X_2 / X_1)\,(1 - k) / k \leq (\hat{x}_1 - X_1) / X_1 \qquad (0.2b)$$

überschätzen, wie es die identischen Umformungen für einen offenen Tabellenwert X zeigen: $X_1 + X_2 \leq k\,X \leftrightarrow$

$$X_1 + X_2 \leq k\,(\hat{x}_1 + X_2) \leftrightarrow (X_1 + X_2) / k - (X_1 + X_2) \leq \hat{x}_1 - X_1 \leftrightarrow (1 + X_2 / X_1)\,(1/k - 1) \leq (\hat{x}_1 - X_1) / X_1.$$

Demnach ist obige „Mindest-Überschätzung" M(2, k) nicht wie die Mindest-Überschätzung M(1, k) im Falle der (1, k)-Dominanzregel für alle Tabellenwerte gleichgroß, sondern hängt über den Faktor $(1 + X_2 / X_1)$ vom Verhält-

nis der beiden größten Einzelangaben ab. Für sehr kleine Wert von X_2 geht dieser Faktor gegen 1, für große X_2 verdoppelt sich der Faktor. Bei der (2, k)-Dominanzregel überdeckt die Mindest-Überschätzung also einen Wertebereich von der einfachen bis zur zweifachen Mindest-Überschätzung der (1, k)-Dominanzregel:

$$(1 - k) / k \; < \; M(2, k) \; \leq \; 2 \, (1 - k) / k \qquad (0.2c)$$

In o.g. Beispiel hätte man den Tabellenwert 100 E im Falle der (2,85)-Dominanzregel wegen $X_1 + X_2 = 85 \text{ E} + 10 \text{ E} > 0{,}85 * 100 \text{ E} = k \, X$ primär geheim zu halten, während der selbe Tabellenwert nach der (1,85)-Dominanzregel offen bliebe. Um diesen Tabellenwert auch bei Anwendung einer (2, k)-Dominanzregel offen zu lassen, muss der Parameter k mindestens 95% gewählt werden. Bei k = 95% liegt die Mindest-Überschätzung der (2, k)-Dominanzregel für alle offenen Tabellenwerte gemäß (0.2c) zwischen $(1 - 0{,}95) / 0{,}95 = 0{,}053$ und $2 * (1 - 0{,}95) / 0{,}95 = 0{,}1053$.

Man sieht, dass bei gleichen Parameterwerten k für die (1, k)- und die (2, k)-Dominanzregel die (2, k)-Dominanzregel bereits Werte sperrt, die nach der (1, k)-Dominanzregel noch offen geblieben wären. Um durch die beiden Regeln etwa gleich viele Primärsperrungen zu verursachen, wählt man den Parameter k der (2, k)-Dominanzregel im Allgemeinen größer als den der (1, k)-Regel. Aber auch dann sind beide Regeln nicht gleichwertig: Tabellenwerte mit sehr großen X_1 und kleinen X_2 werden im Falle $k_2 > k_1$ mit der $(1, k_1)$-Regel eher gesperrt als mit der $(2, k_2)$-Dominanzregel, umgekehrt verhält es sich bei annähernd gleichgroßen X_1, X_2, wo die $(2, k_2)$-Dominanzregel öfter sperrt als die $(1, k_1)$-Regel (vgl. Abschnitt 0.3).

0.2.1.3 (n,k)-Dominanzregeln

Nach obigem Vorbild lässt sich eine (n,k)-Dominanzregel für $1 \leq n < N$ aufstellen: Nach dieser Regel ist ein Tabellenwert mit N Einzelangaben primär geheim zu halten, wenn der Anteil der Summe aus den n größten Einzelangaben $X_1, X_2, ..., X_n$ am Tabellenwert X größer als ein vorzugebender Parameter k ist:

$$X_1 + X_2 + ... + X_n > k \, X \qquad (0.3)$$

Nach der (n,k)-Dominanzregel kann ein Tabellenwert also offen bleiben, wenn die n größten Einzelbeiträge in ihrer Summe höchstens k Prozent des gesamten Tabellenwertes ausmachen.

Mit dieser Regel kann man nun auch Vorwissen über mehr als zwei Einzelbeiträge berücksichtigen. Allgemein gilt: Wenn man unterstellen muss, dass ein „Datenangreifer“ n – 1 Einzelbeiträge eines Tabellenwertes oder deren Summenwert kennt, so ist die primäre Geheimhaltung dieses Tabellenwertes nach der (n,k)-Dominanzregel durchzuführen, um die näherungsweise Aufdeckung eines Einzelbeitrages in jedem Fall zu verhindern.

Eine (n,k)-Dominanzregel entspricht einer Fallzahlregel, die bei n Berichtenden zu einer Primärsperrung führt. Beide Regeln unterscheiden sich nur in Bezug auf das Restaggregat aller nicht unter den n herausgehobenen Werten eingeordneten Beiträge. Dem gemäß geht die (n,k)-Dominanzregel in die ihr entsprechende Fallzahlregel über, wenn die Summe der $N – n$ restlichen Beiträge eines Tabellenwertes mit N Berichtenden gegen Null strebt. Es ist

eben genau der Beitrag der N – n restlichen Werte eines Aggregates, der bei Anwendung der (n,k)-Dominanzregel seinen Schutz gegen eine näherungsweise und nicht nur exakte Offenlegung ausmacht.

0.2.2 Die p%-Regel

Ein Tabellenwert ist nach der sogenannten p%-Regel primär geheim zu halten, wenn der Berichtende des zweitgrößten Beitrags mit seinem Wert X_2 und dem Tabellenwert X den größten Beitrag X_1 als Differenz beider Werte, $X - X_2$, genauer schätzen kann, als es der Parameter p der p%-Regel erlaubt, d.h. wenn gilt

$$(X - X_2) - X_1 < p\, X_1 \qquad (0.4).$$

Ein Tabellenwert ist also nach der p%-Regel primär geheim zu halten, wenn das Restaggregat, der Tabellenwert ohne die beiden größten Einzelwerte, kleiner als der p%-Anteil des größten Einzelwertes ist.

Die p%-Regel wird durch die hohe Genauigkeit der Schätzung des größten Einzelwertes nahegelegt, die der Berichtende mit dem zweitgrößten Einzelwert angeben kann. Diese Motivation haben die p%-Regel und die (2,k)-Dominanzregel gemeinsam. Dennoch gibt es einen gravierenden Unterschied zwischen beiden Regeln: Das Restaggregat, mit dessen Hilfe letztlich beurteilt werden soll, ob ein Tabellenwert primär geheim zu halten ist oder nicht, wird im Falle der p%-Regel auf den größten Einzelwert bezogen, während der Bezugswert bei allen Dominanzregeln immer der Tabellenwert selbst ist.

Der Unterschied in den Bezugsgrößen verhindert, dass man die Parameter beider Auswahlkriterien exakt ineinander umrechnen kann: Die Mindest-Überschätzung des größten Einzelwertes durch den Berichtenden des zweitgrößten hängt im Falle der (2,k)-Dominanzregel gemäß (0.2b) und (0.2c) vom Verhältnis des zweitgrößten zum größten Einzelwert ab, bei der p%-Regel ist sie dagegen konstant. Daher kann man die Ergebnisse beider Kriterien nicht miteinander vergleichen! Ein Vorzug der p%-Regel ist sicherlich, dass sie sich mit dem Intervallschutz von Einzelangaben begründen lässt, was in Abschnitt 4.2.3 mit Hilfe des Quaderverfahrens gezeigt wird.

Im o.g. Beispiel beträgt das Restaggregat ohne die beiden größten Einzelwerte 5 Einheiten: $X - X_1 - X_2 = 100$ E – 85 E – 10 E = 5 E . Bezieht man diese auf den größten Einzelwert, erhält man $(X - X_1 - X_2) / X_1 = 5$ E / 85 E = 5,88 %. Demnach ist der Tabellenwert nach der p%-Regel als primär geheim einzustufen, wenn der vorgegebene Parameter p der p%-Regel etwa 5,89 % oder mehr beträgt.

Im Fall des ganz speziellen Tabellenwertes des Beispiels kann man beide Regeln miteinander vergleichen und beispielsweise den Parameter k bestimmen, der den größten Einzelwert mit einer (2,k)-Dominanzregel genau so gut schützt wie die p%-Regel mit einem Parameter p = 5,89 % . Für den nach der p%- und nach der (2,k)-Regel gerade noch offenen Wert muss dann $(1 + 10 \text{ E} / 85 \text{ E})(1- k) / k \approx 0{,}0588$ gelten. Man sieht, dass die p%-Regel mit p = 5,89 % den Beispielwert etwa so gut schützt wie die (2, 95)-Dominanzregel.

Dies mag Anregung für die Parameterwahl von p bei Orientierung an k sein bzw. auch umgekehrt für die Wahl von k nach Vorgabe von p, wenn der zweite dominierende Einzelbeitrag klein gegen den ersten ist. Bei fast gleichgroßen dominierenden Werten erhält man für den tabellenwertspezifischen Faktor von (0.2b) annähernd 2. Um auch in diesem speziellen Falle mit beiden Regeln gleiche Wirkungen zu erzielen, hat man wieder die Mindest-Überschätzungen beider Regeln zu vergleichen und erhält das Ergebnis, dass die p%-Regel mit p = 5,89% ebenso gut schützt wie die (2, 97)-Dominanzregel.

Beim Vergleich der p%-Regel mit der (1, k)-Dominanzregel lassen sich die Parameter beider Regeln nach dem Kriterium der Mindest-Überschätzung für alle Tabellenwerte einheitlich ineinander umrechnen. Orientiert man sich in obigem Beispiel an der (1, 85)-Dominanzregel, so erhält man für den Parameter der p%-Regel p = 17,6%. Erfahrungsgemäß führt das aber zu erheblich mehr Primärsperrungen als bei der (1, 85)-Dominanzregel, weil die p%-Regel zwar bei sehr kleinen X_2 die gleichen Tabellenwerte sperrt wie die (1, 85)-Dominanzregel, darüber hinaus aber auch bei größeren X_2 noch einen hinreichenden Schutz gegen zu genaues Rückrechnen bietet, wodurch die Anzahl der Primärsperrungen erhöht wird. Eine eingehendere Analyse folgt in Abschnitt 0.3 .

Für den praktischen Einsatz von Primärsperrkriterien ist interessant, dass die p%-Regel – wie im Übrigen auch die (2, k)-Dominanzregel - die Fallzahlregel in Gestalt der Dreier-Regel gleich mitbedienen kann: Ist ein Tabellenwert nach der Dreier-Regel primär geheim, so auch immer nach der p%-Regel. Umfasst der Tabellenwert nämlich weniger als drei Einzelangaben, so ist das Restaggregat nach Abzug der beiden größten Einzelwerte gleich Null und damit auch die linke Seite der Ungleichung (0.4). Ein Tabellenwert mit weniger als drei Berichtenden ist für jeden positiven Parameter p der p%-Regel als primär geheim auszuweisen; entsprechendes gilt für jedes k mit $0 \leq k < 1$ einer (2, k)-Dominanzregel, wenn man dort n = N = 2 zulässt.

0.2.3 Die (p;q)-Regeln

Wen man bei dem zu unterstellenden Vorwissen meint, dass Brancheninsider fremde Einzelangaben bis auf Schätzintervalle genau kennen, so wird man bei der Schutzfehlerabschätzung vom Tabellenwert außer den beiden größten Einzelwerten - in konsequenter Verallgemeinerung der p%-Regel - auch noch den untersten Schätzwert des Restaggregats subtrahieren. Als praktikable Annahme wird nun unterstellt, dass Brancheninsider **jeden** Einzelbeitrag bis auf ±q % genau kennen.

Nach der (p;q)-Regel ist dann ein Tabellenwert primär geheim zu halten, wenn der Berichtende der zweitgrößten Einzelangabe, den größten Einzelbeitrag um weniger als p% überschätzt, indem er seinen eigenen Beitrag, sowie die von ihm auf q% genau geschätzte Mindestgröße der übrigen Beiträge vom Tabellenwert abzieht.

Das bei Berücksichtigung von Schätzintervallen als Vorwissen anzusetzende Auswahlkriterium für die primär geheimzuhaltenden Tabellenwerte lautet zunächst ganz allgemein:

$$X - X_2 - \min\left(\sum_{i=3}^{N} \hat{x}_i\right) - X_1 < p\, X_1 \qquad (0.5a)$$

Die auf der linken Seite der Ungleichung stehende Differenz aus dem Restaggregat $X - X_2 - X_1$ und dem untersten Schätzwert davon, $\min(\sum_{i=3}^{N} \hat{x}_i)$, ist der Schätzfehler des Restaggregats und nach obiger Voraussetzung

$$X - X_2 - X_1 - \min(\sum_{i=3}^{N} \hat{x}_i) = q \sum_{i=3}^{N} x_i = q\,(X - X_2 - X_1)\,.$$

Ein Tabellenwert wird folglich nach der (p,q)-Regel primär geheimgehalten, wenn q% des Restaggregats, des Tabellenwertes ohne die beiden größten Einzelwerte, kleiner als p% des größten Einzelwertes ist, d.h. wenn der Schätzfehler des Restaggregats kleiner als der vorgegebene Schutzfehler des größten Einzelwertes ist:

$$q\,(X - X_2 - X_1) < pX_1 \qquad (0.5b)$$

Aus der letzten Ungleichung folgt, dass X primär geheim zu halten ist, falls $(X - X_1 - X_2) / X_1 < p / q = p'$, falls also die p%-Regel mit dem einzigen Parameter $p' = p / q$ gilt. Die (p,q)-Regel ist daher einer p%-Regel mit größerem p äquivalent und damit verzichtbar, wenn man nur die p%-Regel mit einem hinreichend großen p-Parameter ausstattet! Außerdem entspricht jede p%-Regel einer (p;q)-Regel mit q = 1, wenn also dem Berichtenden mit dem zweitgrößten Einzelbeitrag kein Vorwissen in Gestalt von Schätzintervallen über die übrigen Einzelbeiträge unterstellt werden kann. D.h. man geht davon aus, dass die größte untere Schranke für diese Werte Null ist.

0.3 Darstellung der gebräuchlichsten Konzentrationsmaße

Unter dem Begriff „Konzentrationsmaße“ werden hier alle bisher in die deutsche amtliche Statistik eingeführten Maße zur Beschreibung der Wertekonzentrierung auf nur wenige besonders große Einzelangaben in einem Tabellenwert bezeichnet. Dazu gehören auch intervallbasierte Beschreibungen von Wertehäufungen wie die p%-Regel, die im Folgenden einander gegenübergestellt werden.

Im Folgenden werden ausschließlich n-dimensionale Tabellen mit nicht negativen Werten, sogenannte positive Tabellen, behandelt, in denen keine Nullen gesperrt werden, weder primär noch sekundär. Jeder Tabellenwert X der Veröffentlichungstabelle lässt sich als Summe der zugehörigen Werte X_i, i = 1,2,---, r schreiben, wo r die größte Anzahl von Einzelwerten bezeichnet, die zu einem Aggregat der Veröffentlichungstabelle beitragen.

$$X = \sum_{i=1}^{r} X_i$$

Alle hier auftretenden Indizes beziehen sich ausschließlich auf die Gliederung nach den Einzeldaten. Falls die Anzahl N der Berichtenden eines Tabellenwertes weniger als r Fälle umfasst, wenn also N < r ist, denke man sich dieses Tabellenfeld mit $r - N$ Nullen aufgefüllt: D.h. auch bei der Summenbildung werden nicht vorhandene

Werte X_i durch Nullen ersetzt. Im Folgenden wird außerdem angenommen, dass das Einzeldatenmaterial der Größe nach angeordnet ist, d.h.

$$X_1 \geq X_2 \geq \text{---} \geq X_r \quad .$$

Um die unterschiedlichen Wirkungen der in der amtlichen Statistik bisher verwendeten Konzentrationsmaße, der (1, k)- und der (2, k)-Regel sowie der p%-Regel, zu verdeutlichen, trägt man zweckmäßig alle Werte der betrachteten Veröffentlichungstabelle als Paare der Anteilswerte ihrer beiden größten Einzelwerte in einem x-y-Koordinatensystem ein (M. Walter, StaLa Bayern, AK. mathematische Methodik des Statistischen Bundesamtes am 27.10.2000). $x = X_1 / X$ symbolisiert den größten und $y = X_2 / X$ den zweitgrößten Anteilswert von X. Durch Zusammenfassung aller Einzelwerte mit Indizes größer als 3 zu einem Rest-Aggregat erhält man aus obiger Summe für jeden Veröffentlichungswert X

$$X = X_1 + X_2 + \text{Rest} \quad \text{mit} \quad \text{Rest} = \sum_{i=3}^{r} X_i \; .$$

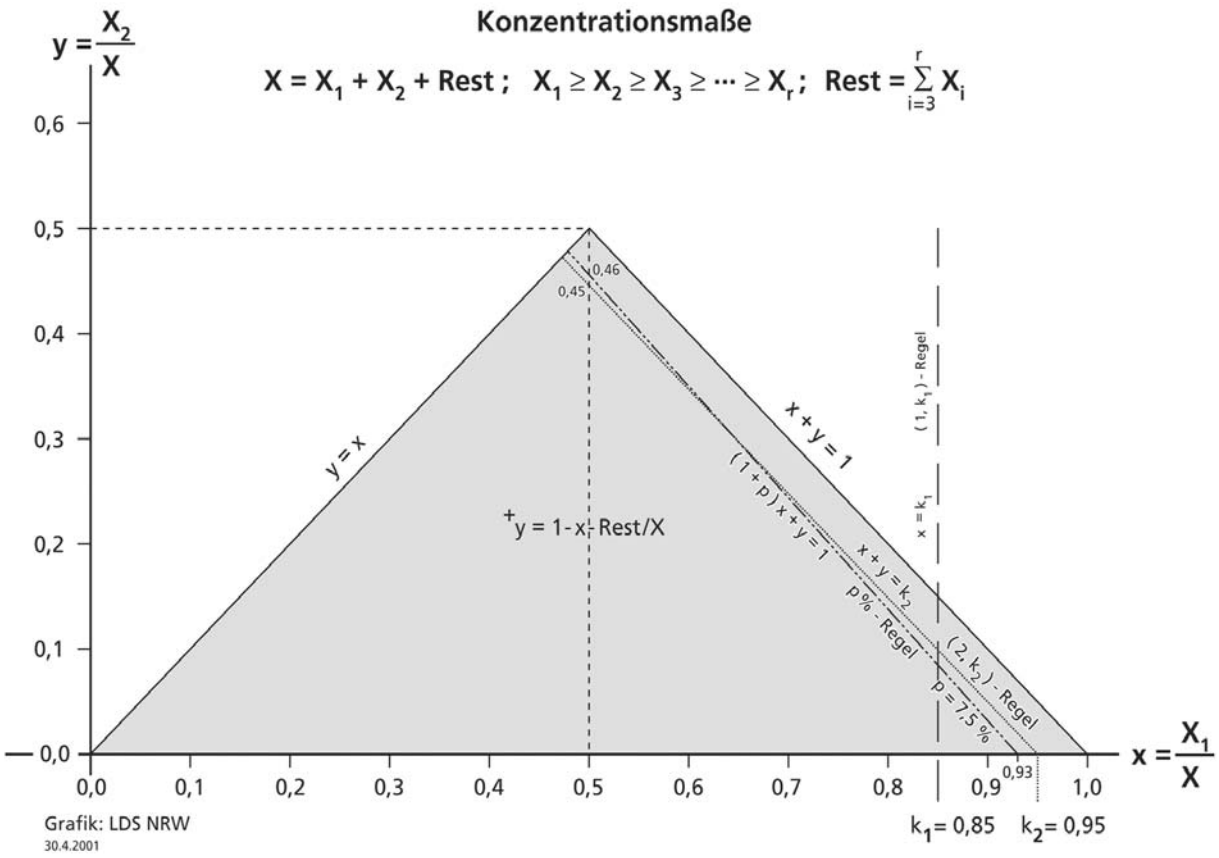
Daraus ergibt sich schließlich ein Zusammenhang zwischen den Koordinatenwerten x und y jedes Veröffentlichungswertes X, indem jeder Wert dieser Summe auf den Aggregatwert X bezogen wird:

$$y = 1 - x - \text{Rest} / X \qquad (0.6)$$

Wegen der vorausgesetzten Positivität der Tabelle liegen alle Aggregate oberhalb der x-Achse ($X_2 > 0$) oder darauf ($X_2 = 0$), wegen der Anordnung der Einzeldaten liegen sie auf der Geraden $y = x$, ($X_1 = X_2$) oder darunter ($X_1 > X_2$) und wegen (0.6) liegen die Aggregate auf der Geraden $y = 1 - x$ (für Rest = 0) oder darunter (für Rest > 0).

Die o.g. Konzentrationsmaße lassen sich bezüglich x, y wie folgt ausdrücken:

(1, k)-Regel:	$X_1 / X > k$	$\leftrightarrow$	$x > k$	(primär geheim)
(2, k)-Regel:	$(X_1 + X_2) / X > k$	$\leftrightarrow$	$x + y > k$	(primär geheim)
p%-Regel:	$X - X_1 - X_2 < p * X_1$	$\leftrightarrow$	$(1 + p)\, x + y > 1$	(primär geheim)



Die hier interessierenden Konzentrationsmaße sind in obigem Diagramm durch Geraden charakterisiert, die die mit Tabellenwerten belegte Dreiecksfläche in jeweils zwei Flächen zerlegen. Davon beherbergt die eine Fläche die primär gesperrten, die andere die nicht primär gesperrten Tabellenwerte. Der jeweilige Bereich nicht primär gesperrter Werte liegt immer links bzw. unterhalb der Geraden der betreffenden Regel oder auch auf der Geraden dieser Regel.

Die Sonderstellung der (1,k)-Regel gegenüber den anderen beiden primären Geheimhaltungsregeln, der (2,k)- und der p%-Regel tritt in dieser graphischen Darstellung deutlich hervor: Mit der (1,k)-Regel wird eben die Dominanz von nur einer Einzelangabe berücksichtigt und nicht die der beiden größten Einzelwerte. Bei geeigneter Parameterwahl unterscheiden sich die beiden anderen Regeln, die (2,k)- und die p%-Regel, nur marginal voneinander. Die in der amtlichen Statistik auch noch diskutierte (p,q)-Regel wurde in diese Betrachtung nicht explizit mit einbezogen, weil sie ohnehin einer p´%-Regel mit einzigem Parameter p´ = p / q äquivalent ist und somit bereits implizit in der Darstellung auftritt (vgl. 0.2.3).

1. Grundlegendes zur sekundären Geheimhaltung

1.1 Prinzipien der sekundären Geheimhaltung

Besonders einfach gestaltet sich die Beschreibung der Sicherung primär geheimer Tabellenwerte bei zweidimensionalen Tabellen, die nicht durch Zwischensummen unterteilt sind, bei denen also in jeder Gliederung nur eine Randsumme auftritt. Mit Hilfe von Beispieltabellen dieser Art werden zunächst die grundlegenden Möglichkeiten und Probleme der sekundären Geheimhaltung erläutert.

1.1.1 Sekundäre Geheimhaltung mit Hilfe von Summensperrungen

Die am einfachsten zu realisierende Vorschrift zur Sicherung geheimer Tabellenwerte fordert, dass Randsummen, zu denen nur ein geheimer Wert beiträgt, nicht veröffentlicht werden dürfen, weil man sonst z.B. die in Abbildung 1.1 durch „●“ markierten primär geheimen Werte einfach durch Differenzbildung aus dem betreffenden Zeilen- oder Spalten-Summenwert und den anderen noch offenen Werten der Zeile oder Spalte des primär geheimen Wertes berechnen könnte.

Abb. 1.1

Kreise	**Gruppe**				
	A	**B**	**C**	**D**	**Σ**
1	11 7760	8 240	4 57	117 4154	140 12211
2	3 240	● ~~2~~ ~~187~~	33 184	67 1782	⊙ ~~105~~ ~~2393~~
3	322 1723	3 316	18 115	● ~~8~~ ~~258~~	⊙ ~~351~~ ~~2412~~
4	116 842	87 448	21 439	4 86	228 1815
Reg.-Bez.	452 10565	⊙ ~~100~~ ~~1191~~	76 795	⊙ ~~196~~ ~~6280~~	824 18831

obere Zeile: Anzahl
untere Zeile: Betrag
● = geheim zu haltender Wert
⊙ = Löschung zur Vermeidung der Errechenbarkeit der geheim zu haltenden Werte

1.1.2 Zielfunktion „Minimale gesperrte Wertesumme“

Randsummensperrungen bedeuten nicht nur einen hohen Informationsverlust; bei Einbindung der betreffenden Tabelle in eine umfassendere hierarchisch gegliederte Gesamttabelle werden u.U. noch weitere Sekundärsperrungen in anderen Teilen der Gesamttabelle erforderlich, die es zu vermeiden gilt (vergleiche Unterpunkt 1.2). Um solche

Randsummen für die Veröffentlichung zurückzugewinnen, kann man die als geheim vorgegebenen, primär geheimen Werte durch Sekundärsperrungen so zu sichern trachten, dass die Summe gesperrter Werte möglichst klein ausfällt.

Dabei muss man berücksichtigen, dass die Sperrung zusätzlicher Werte in der Zeile und Spalte des primär geheimen Wertes in der Regel nicht ausreicht, um einen hinreichenden Schutz gegen eindeutiges Rückrechnen zu gewährleisten. Es muss sichergestellt werden, dass auch die sekundär geheimen Werte nicht berechnet werden können. Dazu sind häufig weitere Sperrungen erforderlich. Als besonders einfaches Sperrverfahren, das einen hinreichenden Schutz gegen Rückrechnung geheimer Werte garantiert, bietet sich bei 2-dimensionalen Tabellen die Karree-Sicherung an, wobei jedem primär geheimen Wert - unabhängig von möglichen anderen geheimen Tabellenwerten - ein Karree mit lauter gesperrten Werten zugeordnet wird.

Ein Karree bezeichnet eine Gesamtheit von vier Tabellenwerten, die als Eckwerte in der Ebene einer zweidimensionalen Tabelle ein geometrisches Rechteck beschreiben. Offensichtlich ist dies die kleinste Gesamtheit von Tabellenwerten, die zur Sicherung nur eines geheimen Wertes herangezogen werden kann: Zur Sicherung gegen Rückrechnung mit der Summenbeziehung der Zeile genügt ein weiterer Tabellenwert als Partner. Zur Sicherung des geheimen Wertes und seines Zeilenpartners müssen bezüglich der Summenbeziehungen in den beiden Spalten zwei weitere Partner ausgewählt werden. Diese schützen sich nur dann wieder gegenseitig bezüglich der Summenbeziehungen in den Zeilen, wenn sie beide in der gleichen Zeile liegen, wenn also alle vier Werte ein Karree beschreiben. Anderenfalls wären weitere Schutzpartner aufzusuchen, womit sich die Schutzpartneranzahl erhöhen würde.

Im Hinblick auf die vorzunehmende Verallgemeinerung des Verfahrens auf n-dimensionale Tabellen sollen solche Karrees zum Schutze geheimer Werte in 2-dimensionalen Tabellen auch als (zweidimensionale) Quader bezeichnet werden, und ein Verfahren, das einen geheimen Wert mit Hilfe eines Quaders zu schützen vermag, als Quaderverfahren.

Es ergibt sich z.B. für das primär geheime Feld (2,B) das Karree {(1,B), (1,C), (2,B), (2,C)} mit einer besonders kleinen Summe zusätzlich zu sperrender Werte: 240+57+184 = 481. Eine nach diesen Kriterien gesicherte Tabelle zeigt die Abbildung 1.2.

Abb. 1.2

Kreise	**Gruppe**				
	A	**B**	**C**	**D**	**Σ**
1	11 7760	⊙ ~~8~~ ~~240~~	⊙ ~~4~~ ~~57~~	117 4154	140 12211
2	3 240	● ~~2~~ ~~187~~	⊙ ~~33~~ ~~184~~	67 1782	105 2393
3	322 1723	3 316	⊙ ~~18~~ ~~115~~	● ~~8~~ ~~258~~	351 2412
4	116 842	87 448	⊙ ~~21~~ ~~439~~	⊙ ~~4~~ ~~86~~	228 1815
Reg.-Bez.	452 10565	100 1191	76 795	196 6280	824 18831

obere Zeile: Anzahl
untere Zeile: Betrag

● = geheim zu haltender Wert
⊙ = Löschung zur Vermeidung der Errechenbarkeit der geheim zu haltenden Werte

Der ersichtliche Nachteil dieses Vorgehens liegt in der unter Umständen großen Anzahl von Sekundärsperrungen.

1.1.3 Zielfunktion „Minimale Anzahl Sekundärsperrungen“

Eine wesentlich kleinere Anzahl von Sekundärsperrungen lässt sich erreichen, wenn man bei der Auswahl von Sicherungskarrees bzw. zweidimensionalen Quadern solche mit bereits gesperrten Tabellenfeldern besonders bevorzugt:

Abb. 1.3

Kreise	**Gruppe**				
	A	**B**	**C**	**D**	**∑**
1	11 7760	8 240	4 57	117 4154	140 12211
2	3 240	● ~~2~~ ~~187~~	33 184	⊙ ~~67~~ ~~1782~~	105 2393
3	322 1723	⊙ ~~3~~ ~~316~~	18 115	● ~~8~~ ~~258~~	351 2412
4	116 842	87 448	21 439	4 86	228 1815
Reg.-Bez.	452 10565	100 1191	76 795	196 6280	824 18831

obere Zeile: Anzahl
untere Zeile: Betrag

● = geheim zu haltender Wert
⊙ = Löschung zur Vermeidung der Errechenbarkeit der geheim zu haltenden Werte

Die Tabelle der Abb. 1.3 zeigt nun den gegenwärtig benutzten Ansatz: Der zur Sicherung eines (beliebigen) geheimen Wertes aufzusuchende Quader ist so auszuwählen, dass er möglichst viele bereits gesperrte Werte enthält, dass also möglichst wenige noch offene Quaderwerte zur Sicherung des betrachteten geheimen Wertes zusätzlich gesperrt werden müssen. Erst in zweiter Linie, d.h. wenn mehrere Quader mit der selben Anzahl noch zu sperrender Werte zur Auswahl stehen, soll deren Wertesumme (der noch offenen Werte) minimal sein (siehe auch „EDV-Verfahren zur Wahrung der Geheimhaltung...“, Statistische Rundschau NRW 1991).

1.2 Durch Zwischensummen untergliederte Tabellen

1.2.1 Begründung einer Untertabellenhierarchie

Tabellen, deren Werte und Merkmalsträgerzahlen wie in den oben angeführten Beispielen bezüglich jedes Gliederungskriteriums zu nur einer (Rand-)Summe aufaddiert werden, sind in einer mehrfach durch Zwischensummen untergliederten Gesamttabelle nur als Teilgesamtheiten realisiert. Solche Teilgesamtheiten werden im Folgenden als Untertabellen bezeichnet.

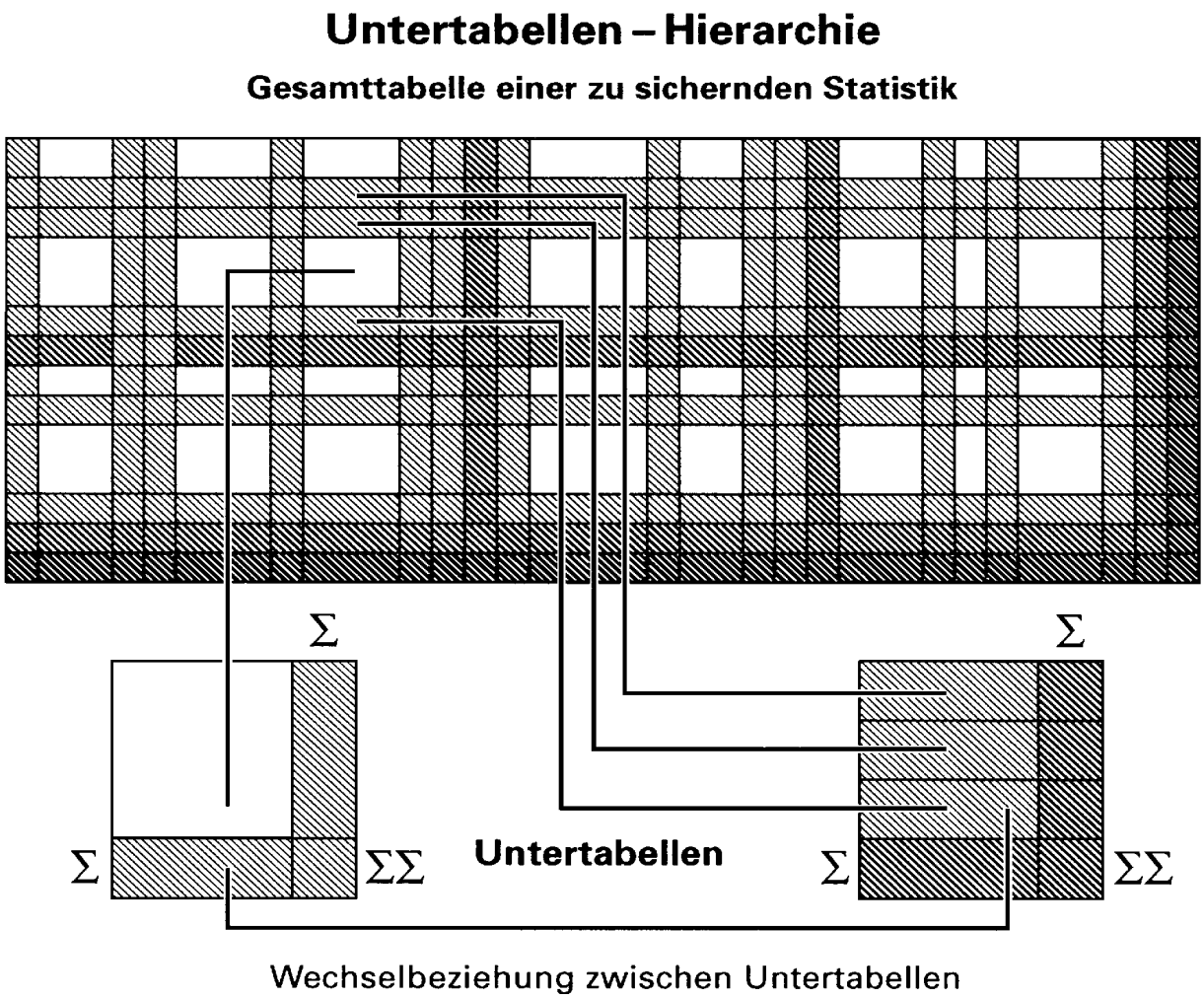
Definition:

Eine (n-dimensionale) Untertabelle ist eine Teilgesamtheit einer durch Zwischensummen unterteilten n-dimensionalen (Gesamt-) Tabelle, die nach den selben n Merkmalen gegliedert ist wie die Gesamttabelle, die aber keine Zwischensummen, sondern in jeder Gliederung nur eine Randsumme ausweist.

Im Falle der eingangs aufgeführten zweidimensionalen „Eingabedaten“ (unterer Teil der Abbildung Eingabedaten) erhält man z.B. eine Untertabelle, indem in sachlicher Gliederung die nur zu einem Dreisteller aufaggregierten Viersteller und in regionaler Gliederung die nur zu einem Kreis beitragenden Gemeinden nebst ihren Summenfeldern, dem zugehörigen Dreisteller und dem zugehörigen Kreis in eine Tabelle aufgenommen werden. Auf diese Weise lassen sich auch höher aggregierte Untertabellen extrahieren, indem z.B. die Dreisteller mit zugehörigem Zweisteller und die Gemeinden mit zugehörigem Kreis in einer Untertabelle zusammengefasst werden. Zwei Untertabellen dieser Art sind in der Abbildung 1.4 symbolisch dargestellt. Für eine eingehendere Betrachtung der Untertabellenaufstellung und -organisation wird auf den o.g. Beitrag in „Statistische Rundschau NRW“ und besonders auf „Safeguarding Secrecy in Aggregative Data“, Dublin 1992 verwiesen.

Durch die Behandlung von einzelnen Untertabellen wird nun das Problem der Geheimhaltung der Gesamttabelle auf ganz natürliche Weise in eine Vielzahl kleiner überschaubarer Teilprobleme zerlegt (siehe die schematische Darstellung):

Abb. 1.4



Ähnlich wie die Geheimhaltung eines einzelnen Wertes durch seine Einbindung in eine Tabelle "gefährdet" wird, verhält es sich mit ganzen Untertabellen, die in die Aggregationsstufenhierarchie einer Gesamttabelle eingeordnet sind. Wie sich solche Untertabellen bei der Wahrung der Geheimhaltung gegenseitig beeinflussen können, zeigen die Beispieltabellen (Abb. 1.5, Abb. 1.6) in Verbindung mit der schematischen Darstellung (Abb. 1.4):

Bei schwach besetzten Tabellen kann es sein, dass sich mit den Werten gleicher Aggregationsstufen kein "Karree" für die Sicherung eines geheimen Wertes finden lässt. Hier muss man auf Summenwerte ausweichen. So entstehen neue geheime Werte in einer Tabelle der nächsthöheren Aggregationsstufe, die dann in dieser Tabelle gesichert werden müssen!

Das primär geheime Feld (2,B) der Beispieltabelle Abb. 1.5 lässt sich zwar durch Sperren von (2,C) gegen Rückrechnung durch Differenzbildung in der Zeile 2 schützen, bezüglich der Spalten fehlen aber besetzte sperrbare Tabellenfelder: Die Sekundärsperrung (2,C) kann nur durch die Summensperrung (Reg.-Bez., C) gesichert werden. Daraus ergibt sich das Karree {(2,B), (2,C), (Reg.-Bez., B), (Reg.-Bez., C)} mit zwei erzwungenen Randsummensperrungen.

Abb. 1.5

Kreise	**Gruppe**				
	A	**B**	**C**	**D**	**∑**
1		8 240			8 240
2	3 240	● ~~2~~ ~~187~~	⊙ ~~33~~ ~~184~~	67 1782	105 2393
3		3 316			3 316
4		87 448			87 448
Reg.-Bez.	3 240	⊕ ~~100~~ ~~1191~~	⊕ ~~33~~ ~~184~~	67 1782	203 3397

obere Zeile: Anzahl
untere Zeile: Betrag

● = geheim zu haltender Wert
⊕ = Leere Tabellenfelder erzwingen Summensperrungen
⊙ = Löschung zur Vermeidung der Errechenbarkeit der geheim zu haltenden Werte

Im Laufe des Sicherungsvorganges der Gesamttabelle können Sekundärsperrungen in Tabellen höherer Aggregation auftreten; diese findet man in den zugehörigen Tabellen niedrigerer Verdichtung als Summensperrungen wieder. Hier sind unter Umständen zusätzliche Sperrungen im Inneren der Tabelle nötig.

Abb. 1.6

Kreise	**Gruppe**				
	A	**B**	**C**	**D**	**∑**
1	11 7760	8 240	4 57	117 4154	140 12211
2	⊕ ~~3~~ ~~240~~	● ~~2~~ ~~187~~	⊕ ~~33~~ ~~184~~	⊙ ~~67~~ ~~1782~~	105 2393
3	322 1723	⊙ ~~3~~ ~~316~~	18 115	● ~~8~~ ~~258~~	351 2412
4	116 842	87 448	21 439	4 86	228 1815
Reg.-Bez.	ι ~~452~~ ~~10565~~	100 1191	ι ~~76~~ ~~795~~	196 6280	824 18831

obere Zeile: Anzahl
untere Zeile: Betrag

● = geheim zu haltender Wert
⊙ = Löschung zur Vermeidung der Errechenbarkeit der geheim zu haltenden Werte
ι = Löschung höherer Hierarchiestufen
⊕ = Löschung in höherer Hierarchiestufe erzwingt Sperrung

Die Sicherung jeder Untertabelle für sich alleine betrachtet, d.h. herausgelöst aus ihrer Untertabellenhierarchie, stellt eine aus der Sicht der Gesamttabelle i.A. unzulässige Idealisierung dar, weil Sperrungen in Untertabellen

höherer Aggregationsstufen immer auch Sperrungen in den zugehörigen Untertabellen niedrigerer Verdichtung bedeuten. Nur dann, wenn sich ausnahmsweise aufgrund günstiger Tabellenfeldbelegungen alle Sperrungen, primäre wie sekundäre, in jeder Gliederung auf das unterste Niveau beschränken, können die Untertabellen unabhängig voneinander gesichert werden, in allen anderen Fällen sind sie in Bezug auf die Geheimhaltung voneinander abhängig.

Das hier zunächst anhand von zweidimensionalen Beispieltabellen eingeführte Quaderkonzept zur Sicherung geheimer Werte bezieht sich demgegenüber immer nur auf Tabellen, die nicht durch Zwischensummen unterteilt sind.

Aus diesem Grunde bietet sich ein zweistufiges heuristisches Verfahren an: Die erste Stufe sichert mit dem Quaderverfahren die Geheimhaltung in jeder einzelnen Untertabelle, die zweite Stufe umfasst den gegenseitigen Abgleich aller Untertabellen.

Es sei bereits an dieser Stelle angemerkt, dass auch noch eine andere Möglichkeit besteht, eine mehrfach durch Zwischensummen unterteilte Tabelle so zu organisieren, dass sie mit dem Quaderverfahren bearbeitet werden kann: die bereits in der Einführung erwähnte Aufstockung der Tabellendimension. Diese Möglichkeit wird erst im sechsten Kapitel, das sich mit sogenannten überlappenden Tabellen befasst, eingehend diskutiert.

1.2.2 Beispieltabelle mit Zwischensummen in zwei Gliederungen

Um die wechselseitige Abhängigkeit der Untertabellen einer mehrfach durch Zwischensummen unterteilten Statistiktabelle noch an einem Beispiel zu verdeutlichen, wurde die in Abbildung 1.7 aufgeführte nach einem numerischen und nach einem alphanumerischen Schlüssel gegliederte Tabelle nach Eintrag der Primärsperrungen (gekennzeichnet mit P) mit dem Sekundärsperrverfahren zur Vermeidung einer exakten Rückrechnung der primär geheimen Werte bearbeitet und die Sekundärsperrungen durch S kenntlich gemacht. Dabei wurden Tabellenfelder mit Wert = 0 und Fallzahl = 0 wie leere Tabellenfelder behandelt, d.h. nicht als Sperrkandidaten betrachtet.

Die drei Zwischensummenspalten AC, AB, AA enthalten keine Sperrvermerke. Sie begrenzen daher 3 Spaltenstreifen, die hinsichtlich des Sperrvorganges vollkommen unabhängig voneinander bearbeitet werden können, weil das Gleichungssystem jedes Streifens zur Berechnung der gesperrten Werte keine gesperrten Werte eines anderen Spaltenstreifens enthält. Im Folgenden wird von rechts nach links ein Spaltenstreifen nach dem anderen abgearbeitet.

Am einfachsten gestaltet sich der Sperrvorgang im rechten Spaltenstreifen, bestehend aus den Spalten AAD, AAC, AAB, AAA mit Randsummenspalte AA: Die Sicherung der drei primär geheimen Werte in den Feldern (134; AAA), (113; AAD) und (113; AAB) kann durch Karrees auf den untersten Aggregationsstufen erfolgen; es treten also keine Zwischensummensperrungen auf. Alle Untertabellen des oben bezeichneten rechten Spaltenstreifens sind hinsichtlich des Sperrprozesses unabhängig voneinander, das heißt, wie Einzeltabellen z.B. nach dem zu Abbildung 1.3 angegebenen Vorgehen zu behandeln.

Der mittlere Spaltenstreifen, die Spalten ABC, ABB, ABA mit Zwischensummenspalte AB, liefert ein Beispiel für voneinander abhängige Untertabellen, die aneinander abgeglichen werden müssen. Ihre Abhängigkeit wird verursacht durch den primär geheimen Wert in erster Spalten-, aber zweiter Zeilenaggregation im Tabellenfeld (110; ABA). Außerdem sind in diesen Spaltenstreifen noch zwei weitere Primärsperrvermerke eingetragen, in den Feldern (123; ABB), (112; ABA) auf den niedrigsten Aggregationsstufen. Es erweist sich im Allgemeinen als zweckmäßig, wenn man mit der Sicherung von geheimen Werten höchster Aggregationsstufen beginnt, weil durch Sperrungen in höheren Hierarchiestufen in der Regel weitere Sicherungen in den zugehörigen Untertabellen niedrigerer Verdichtung hinzukommen, die dann bei der Sicherung von geheimen Werten auf diesen unteren Ebenen mitberücksichtigt werden können.

Der am höchsten aggregierte primär geheime Tabellenwert im mittleren Spaltenstreifen befindet sich im Feld (110; ABA). Die zugehörige Untertabelle mit derselben Zeilen- und Spaltenaggregation ist durch die Zeilen 110, 120, 130 und die Summenzeile 100 innerhalb des Mittelstreifens gegeben. Als Karrees im Inneren dieser Untertabelle stehen daher zur Auswahl {(110; ABA), (110; ABB), (130; ABA), (130; ABB)}und {(110; ABA), (110; ABB), (120; ABA), (120; ABB)}. Davon hat das zweite Karree die kleinere Summe zusätzlich zu sperrender Werte; die noch offenen Werte dieses Karrees werden daher mit dem Sperrvermerk S versehen. Die beiden Primärsperrungen niedrigster Aggregation, (123; ABB), (112; ABA) werden dann in ihren Untertabellen, dem Zeilenstreifen 120 bis 125 bzw. 110 bis 113 im Mittelspaltenstreifen unter Zuhilfenahme der bereits gesperrten Randsummenwerte gesichert. Dadurch ergeben sich die Karrees geheimer Tabellenwerte {(120; ABA), (120; ABB), (123; ABA), (123; ABB)} und {(110; ABA), (110; ABB), (112; ABA), (112; ABB)}. Die Primärsperrungen des mittleren Spaltenstreifens sind damit vollständig gegen eindeutige Rückrechnung gesichert.

Der linke Spaltenstreifen mit den Spalten ACD, ACC, ACB, ACA und der Summenspalte AC weist nur Primärsperrungen mit niedrigster Zeilen- und Spaltenaggregation aus und zwar in jeder seiner Untertabellen niedrigster Aggregation: In der obersten Untertabelle, gekennzeichnet durch die Zeilen 130 bis 134 sind das die Werte in den Feldern (134; ACC) und (133; ACD); der mittlere Streifen mit den Zeilen 120 bis 125 enthält ebenfalls zwei primär geheime Werte, die Felder (124; ACC) und (122; ACC); im untersten Zeilenstreifen, der die Zeilen 110 bis 113 überdeckt, findet man nur einen primär geheimen Wert im Tabellenfeld (113; ACD).

Bei der Sicherung des primär geheimen Wertes in der obersten Zeile liegt es nahe, nach dem Muster der Abbildung 1.3 das Karree {(134; ACC), (134; ACD), (133; ACC), (133; ACD)} auszuwählen, um mit nur zwei zusätzlichen Sekundärsperrungen (134; ACD) und (133; ACC) bereits beide primär geheimen Werte in der oberen Untertabelle gegen eindeutige Rückrechnung zu schützen. An dieser Stelle kommt nun ein neuer Aspekt in die Diskussion des Sperrvorgangs, die besondere Bewertung von Einzelangaben: Bei der Sicherung von primär geheimen Werten gegen ihre Rückrechenbarkeit ist zu beachten, dass der einzige Berichtende im Feld (133; ACD) der Einzelangabe seinen Wert genau kennt und somit alle Werte des oben angegebenen Sicherungsquaders berechnen kann. Dieser Quader ist daher für den Schutz der Einzelangabe geeignet, weil nur der zu schützende Einzelmelder allein und kein anderer die Quaderwerte berechnen kann; für den primär geheimen Wert in der obersten Zeile bietet der Quader keinen Schutz, es muss ein Karree ohne Einzelangaben als Schutzpartner ausgewählt werden. Als solches bietet sich das Karree {(134; ACA), (134; ACC), (133; ACA), (133; ACC) } an.

In der Untertabelle des Zeilenmittelstreifens sind zwei Einzelangaben als primär geheime Werte eingetragen; auch hier muss abweichend vom Sperrmuster der Abbildung 1.3 für jeden primär geheimen Wert ein Quader gefunden werden, der außer dem zu schützenden Wert selbst sonst keine weiteren Einzelangaben enthält. Dabei kann man bei der Sicherung der zweiten Einzelangabe auf Sekundärsperrungen, die durch die Sicherung der ersten Einzelangabe verursacht wurden, zurückgreifen, um so die Anzahl der Sekundärsperrungen möglichst klein zu halten. Die beiden zur Sicherung der beiden Einzelangaben des Zeilenmittelstreifens ausgewählten Karrees {(124; ACC), (124; ACD), (125; ACC), (125; ACD)} und {(122; ACC), (122; ACD), (125; ACC), (125; ACD)} haben dem gemäß die beiden Tabellenfelder in der Zeile 125 gemeinsam, ohne einen der beiden Einzelmelder zu befähigen, diese gemeinsamen Werte zu berechnen. Und zwar könnte der eine der beiden Einzelberichtenden seinen Quader und damit auch die Sekundärsperrungen in Zeile 125 aufdecken, wenn er als alleinige Primärsperrung mit nur einem Quader eingetragen wäre, der zweite Einzelmelder, dessen Angabe der erste nicht kennt, verhindert aber die Rückrechnung der Sekundärsperrungen (durch beide Einzelmelder), so dass beide Einzelangaben durch zwei Karrees mit gemeinsamen Sekundärsperrungen vollständig gesichert sind. Diese Art der Sicherung von primär geheimen Werten ist nicht mit der in Abschnitt 2.1 beschriebenen Doppelquadersicherung zu verwechseln! In diesem Fall ist für den Schutz aller in der Beispieltabelle, Abb. 1.7, eingetragenen Primärsperrungen immer nur ein Sicherungsquader aufzusuchen und kein Doppelquader.

Die hohe Schutzbedürftigkeit von Einzelangaben ist insbesondere dann zu berücksichtigen, wenn in der Veröffentlichungstabelle alle Eintragungen über die Anzahl der Berichtenden von den Sperrungen in den Tabellenfeldern nicht betroffen sind, wenn also durchweg alle Fallzahlen veröffentlicht werden. Dann weiß jeder Einzelmelder in der Veröffentlichungstabelle unmittelbar, dass nur er zu seinem Tabellenfeld beiträgt und ist damit als Schutzpartner in einem Sicherungsquader ungeeignet. Werden die Fallzahlen nicht veröffentlicht, die zu geheimen Werten beitragen, so kommt der Annahme erhöhter Schutzbedürftigkeit von Einzelangaben das Eintragen von Vorinformationen über die Tabellenwerte gleich: Man unterstellt, dass dem Einzelmelder bekannt ist, dass nur er in sein Tabellenfeld eingegliedert werden kann. Solche Vorinformationen, zu denen insbesondere das Wissen über eine Veröffentlichungstabelle gehört, dass diese keine negativen Werte beinhaltet, führen in aller Regel zu einer wesentlichen Verschärfung des Geheimhaltungsproblems und damit auch zu mehr Sperrungen. Die Wirkungen von Vorinformationen werden in gesonderten Abschnitten noch eingehend behandelt.

Die im linken Spaltenstreifen durch den untersten Zeilenstreifen festgelegte Untertabelle weist einen primär geheimen Wert aus, der nicht mehr im Inneren dieser Untertabelle niedrigster Aggregation gesichert werden kann. Es liegt die in Abbildung 1.5 dargestellte Situation vor, wo man im Tabelleninneren kein Karree zum betrachteten geheimen zu sichernden Wert findet; es muss ein Karree mit Randsummenwerten ausgewählt werden, hier das Karree {(113; ACB), (113; ACD), (110; ACB), (110; ACD)}. Diese Rücksperrungen in eine höhere Hierarchieebene erzwingen einen Abgleich mit der entsprechenden Untertabelle zweiter Zeilen- und erster Spaltenaggregation: Betroffen ist die Untertabelle mit den Zeilen 110, 120, 130 und der zugehörigen Summenzeile 100 des linken Spaltenstreifens. Zu sichern sind die beiden Sekundärsperrungen in den Feldern (110; ACB) und (110; ACD); dazu stehen die entsprechenden Felder in den Zeilen 120 und 130 in den Spalten ACB und ACD zur Auswahl. Die kleinste Summe zusätzlich zu sperrender Werte weist das Karree {(110; ACB), (110; ACD), (130; ACB), (130; ACD)} aus, so dass die diesem Karree angehörenden Felder in Zeile 130 zusätzlich zu sperren sind.

Durch die Sekundärsperrungen in Zeile 130 tritt der mit Abbildung 1.6 erläuterte Sicherungsfall auf: Sperrungen in höherer Hierarchie erzwingen weitere Sperrungen in derjenigen Untertabelle in der diese gesperrten Werte Randsummenwerte sind, hier in der oberen durch die Zeilen 130 bis 134 festgelegten Untertabelle unterster Aggregation. Um bei der Sicherung der Sekundärsperrungen in Zeile 130 möglichst viele bereits gesperrte Tabellenwerte mit einzubeziehen, das heißt um möglichst wenige Werte zusätzlich sperren zu müssen, kommt bei der Auswahl von Sicherungsquadern hier nur das Karree {(130; ACB), (130; ACD), (134; ACB), (134; ACD)} in Frage mit dem einzigen zusätzlich zu sperrenden Wert im Tabellenfeld (134; ACB). Damit ist auch der linke Spaltenstreifen vollständig gesichert und somit die Sicherung der gesamten Beispieltabelle abgeschlossen.

Beispieltabelle mit Zwischensummen in zwei Gliederungen Abb. 1.7

1. Schlüssel	2. Schlüssel	ACD	ACC	ACB	ACA	AC	ABC	ABB	ABA	AB	AAD	AAC	AAB	AAA	AA	A
	00000134	112 5 S	10 2 P	1.445 20 S	549 12 S	2.116 39	4.128 34	345 15	211 12	4.684 61	321 21 S	0 0	0 0	95 2 P	416 23	7.216 123
	00000133	40 1 P	66 4 S	0 0	23 3 S	129 8	2.567 44	2.332 30	432 21	5.331 95	732 51	644 34	0 0	0 0	1.376 85	6.836 188
	00000132	723 9	254 11	327 5	543 19	1.847 44	1.123 64	4.427 59	1.632 26	7.182 149	432 23	0 0	234 36	0 0	666 59	9.695 252
	00000131	2.156 33	1.342 23	1.111 17	99 4	4.708 77	590 11	2.334 28	342 9	3.266 48	34 3 S	0 0	0 0	256 17 S	290 20	8.264 145
	00000130	3.031 48 S	1.672 40	2.883 42 S	1.214 38	8.800 168	8.408 153	9.438 132	2.617 68	20.463 353	1.519 98	644 34	234 36	351 19	2.748 187	32.011 708
	00000125	321 5 S	11 3 S	411 18	0 0	743 26	0 0	56 5	0 0	56 5	712 50	3.421 84	0 0	0 0	4.133 134	4.932 165
	00000124	56 4 S	12 1 P	2.152 29	399 11	2.619 45	0 0	123 10	0 0	123 10	345 44	2.612 61	55 3	0 0	3.012 108	5.754 163
	00000123	99 8	311 10	754 19	345 16	1.509 53	221 7	34 2 P	73 6 S	328 15	123 23	321 41	567 32	43 4	1.054 100	2.891 168
	00000122	1.837 33 S	19 1 P	88 4	0 0	1.944 38	0 0	621 13	0 0	621 13	1.015 89	2.221 52	96 18	641 8	3.973 167	6.538 218
	00000121	344 15	298 13	0 0	934 9	1.576 37	0 0	74 8	0 0	74 8	0 0	231 33	0 0	875 72	1.106 105	2.756 150
	00000120	2.657 65	651 28	3.405 70	1.678 36	8.391 199	221 7	908 38 S	73 6 S	1.202 51	2.195 206	8.806 271	718 53	1.559 84	13.278 614	22.871 864
	00000113	53 2 P	221 8	29 3 S	1.001 19	1.304 32	0 0	0 0	0 0	0 0	11 2 P	0 0	21 2 P	0 0	32 4	1.336 36
	00000112	423 18	0 0	0 0	0 0	423 18	0 0	261 5 S	34 2 P	295 7	745 71	0 0	67 8	0 0	812 79	1.530 104
	00000111	28 5	0 0	0 0	0 0	28 5	0 0	0 0	0 0	0 0	148 25 S	0 0	81 7 S	0 0	229 32	257 37
	00000110	504 25 S	221 8	29 3 S	1.001 19	1.755 55	0 0	261 5 S	34 2 P	295 7	904 98	0 0	169 17	0 0	1.073 115	3.123 177
	00000100	6.192 138	2.544 76	6.317 115	3.893 93	18.946 422	8.629 160	10.607 175	2.724 76	21.960 411	4.618 402	9.450 305	1.121 106	1.910 103	17.099 916	58.005 1.749

Legende: Wert 10.000 / Berichtspfl. 100 P — Sperrvermerk (P=primär, S=sekundär)

1.2.3 Organisation der Untertabellengesamtheit einer Statistiktabelle

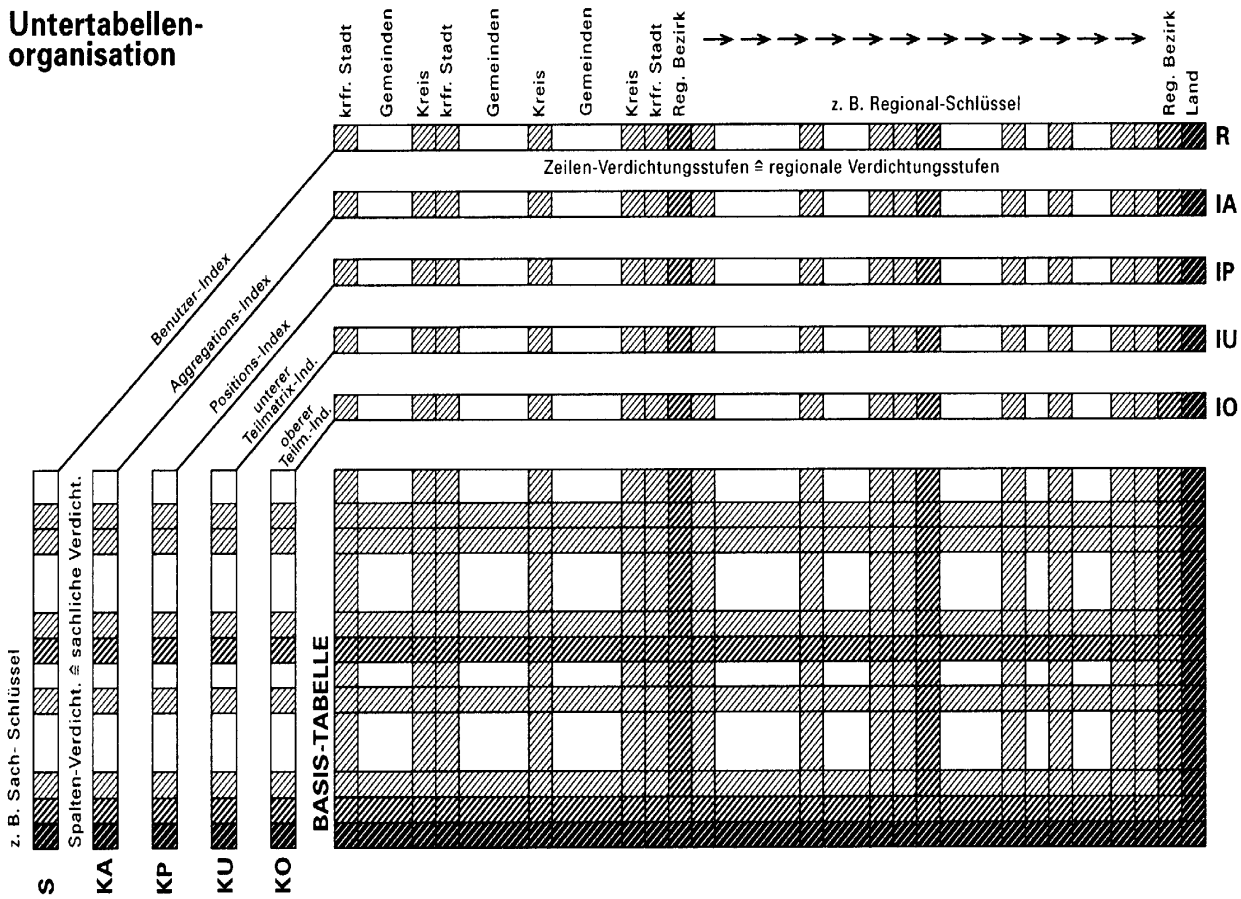
Da jede Sekundärsperrung in einer Untertabelle höherer Verdichtung immer auch eine Summensperrung in einer der zugehörigen Untertabellen niedrigerer Aggregationen bedeutet, wird jeweils mit der Bearbeitung der Untertabellen höchster Aggregationsstufen begonnen und so fortfahrend nach absteigenden Aggregationsstufen, bis alle Untertabellen gesichert sind. Dabei werden die laufend in die Gesamttabelle eingetragenen Sekundärsperrungen der anderen Untertabellen mitberücksichtigt (Untertabellenabgleich).

Dennoch muss das Verfahren erfahrungsgemäß drei- bis viermal durchlaufen werden, weil Summensekundärsperrungen in einer höheren als der gerade bearbeiteten Hierarchiestufe gesichert werden müssen, die beim weiteren Durchlaufen nach absteigenden Aggregationsstufen aber nicht mehr erreicht wird. Das Verfahren iteriert dabei so lange, bis nach einem vollen Durchlauf keine neuen Sperrungen mehr in die Gesamttabelle eingetragen werden müssen.

Um dabei alle Untertabellen einer Statistik in Programmschleifen abarbeiten zu können, muss jede einzelne von ihnen als ganzes ansprechbar sein. Als „Ansprechmerkmale“ sind die Aggregationsstufennummern der Gliederungskriterien einer Untertabelle ohne ihre Randsummen (z.B. die Aggregationsstufen der sachlichen und der regionalen Gliederungsmerkmale der Abb. „Eingabedaten“ unten) zu verwenden, damit die Abarbeitung nach absteigenden Aggregationsniveaus erfolgen kann.

Leider ist die Kennzeichnung einer Untertabelle allein durch diese Aggregationsstufennummern, hier als Aggregationsindizes bezeichnet, nicht eindeutig, so gibt es beispielsweise viele Untertabellen mit Aggregationsstufe 1 für die Zeilen- und Spaltenaggregation wie in Abb. 1.4 links unten. Es muss noch die Lage der auszuwählenden Untertabelle bezüglich aller anderen Untertabellen mit gleichen Aggregationsindizes im Gesamttableau festgelegt werden. Dies geschieht mit Hilfe von Positionsindizes; für jedes Gliederungskriterium einer Untertabelle wird ein Positionsindex angelegt. Eine zweidimensionale Untertabelle ist dem gemäß durch zwei Aggregations- und zwei Positionsindizes eindeutig festgelegt. Diese vier Indizes werden als Ansprechmerkmale einer zweidimensionalen Untertabelle gewählt; sie sind in Abb. 1.8 als Kopfzeile bzw. als Vorspalte angefügt.

Abb. 1.8



Die in Abb. 1.4 links unten dargestellte Untertabelle ist beispielsweise durch die Aggregationsindizes 1;1 bezüglich ihrer Zeilen- und Spaltenaggregation zu beschreiben, die Positionsindizes sind 2;3, weil sie in Bezug auf die 1. Zeilen- bzw. bezüglich der 1. Spaltenaggregationsstufe, von oben bzw. von links gezählt, die zweite bzw. die dritte Position einnimmt. Die rechts unten gezeigte Untertabelle hat die Aggregationsindizes 2 bezüglich der Zeilenaggregation und 1 bezüglich der Spaltenaggregation. Als Positionsindizes hat man 1 bezüglich der Zeilenposition und 3 bezüglich der Spalten (es ist die 1. Untertabelle von oben und die 3. Untertabelle von links mit Aggregationsindizes 2;1 in der Gesamttabelle).

Um eine durch ihre Aggregations- und Positionsindizes adressierte Untertabelle aus der Gesamttabelle Tabellenfeld für Tabellenfeld in einen Arbeitsbereich zu übertragen, sind jedem Indexpaar (Aggregationsindex; Positionsindex) untere und obere Teilmatrixindizes zugeordnet, die angeben, von wo bis wo sich die betreffenden Untertabellenteile innerhalb der Gesamttabelle erstrecken. Bei zusammenhängenden Untertabellen wie im Falle der Beispieltabelle der Abb. 1.4 links unten genügt ein Indexpaar von Teilmatrixindizes für jedes Paar von Aggregations- und Positionsindizes für jedes Gliederungskriterium. Im Falle der rechten unteren Untertabelle der Abb. 1.4 hängen die Zeilen der Untertabelle in der Gesamttabelle nicht zusammen und werden daher einzeln durch ein Teilmatrixindexpaar für jede Zeile festgelegt.

Das hier für zweidimensionale Tabellen dargestellte Konzept zur Wahrung der Geheimhaltung lässt sich direkt auf n-dimensionale mehrfach durch Zwischensummen unterteilte Tabellen verallgemeinern (Dublin, 1992). Insbesondere erfolgt die Sicherung geheimer Werte in einer n-dimensionalen Untertabelle ganz analog zu der an obigen Beispieltabellen erläuterten zweidimensionalen "Karreesicherung" mit Hilfe 2^n Eckwerte umfassender n-dimensionaler Quader. Bei der Organisation n-dimensionaler Untertabellen sind an Stelle der bei zweidimensionaler Gliederung zur Kennzeichnung benutzten Quadrupel aus zwei Aggregations- und zwei Positionsindizes 2n-Tupel zu verwenden, ein Aggregations- und ein Positionsindex für jedes Gliederungskriterium.

1.3 Begründung des Quaderverfahrens

Bei höherdimensionalen Tabellen erweist sich schon allein die Prüfung, ob ein geheimer Tabellenwert bereits gesichert ist oder nicht, als sehr zeitaufwendig. Die Bearbeitung dieser Aufgabe erfordert streng genommen die Aufstellung und Lösung eines linearen Gleichungssystems mit den geheimen Werten als Unbekannte. Eine naheliegende Vereinfachung des Problems bietet sich durch die Reduktion auf unabhängige Einzelgleichungen mit dem Differenzenverfahren an: Es prüft für jede Dimension, ob der geheime Wert der einzige geheime Wert ist, der zu einer Summe beiträgt oder nicht, d.h. es untersucht, ob sich der geheime Wert durch Differenzbildung mit einem Summenwert und den anderen zu dieser Summe beitragenden Werten berechnen lässt oder nicht. Diese besonders Rechenzeit sparende Prüfung ist für die Sicherung des betreffenden geheimen Wertes zwar notwendig, nicht aber hinreichend (siehe dazu die Gegenbeispieltabelle von L.H. COX, 1980).

Abb. 1.9 zeigt so eine Gegenbeispieltabelle mit einer zu isolierenden Zelle (3;C): Die Anzahl der Berichtenden wie auch der von ihnen gemeldete Wert lässt sich für diese Zelle berechnen, indem man von der aus der zweiten und dritten Zeile zusammengefassten Gleichung die beiden Spaltengleichungen der Spalten B und D subtrahiert. Zur Berechnung des Inhaltes der Zelle (3;C) steht gleich noch ein weiteres Gleichungssystem zur Verfügung: Man subtrahiere von der zusammengefassten Spaltengleichung der Spalten A und C die zusammengefasste Zeilengleichung aus der ersten und der vierten Zeile. Unterhalb der Tabelle von Abb. 1.9 wird als Beispiel der Tabellenwert der Zelle (3;C) mit Hilfe des zuerst beschriebenen Gleichungssystems berechnet.

Abb. 1.9

KREISE	Gruppe				
	A	B	C	D	Σ
1	X_1	8 240	X_2	117 4154	140 12211
2	3 240	X_3	33 184	X_4	105 2393
3	322 1723	X_5	X_6	X_7	351 2412
4	X_8	87 448	X_9	4 86	228 1815
Reg.-Bez.	452 10565	100 1191	76 795	196 6280	824 18831

Zeilen :

$$X_3 + X_4 = 2393 - 240 - 184 = 1969$$
$$(X_5 + X_7) + X_6 = 2412 - 1723 = 689$$
$$(X_3 + X_4 + X_5 + X_7) + X_6 = = 2658$$

Spalten :

$$X_3 + X_5 = 1191 - 240 - 448 = 503$$
$$X_4 + X_7 = 6280 - 4154 - 86 = 2040$$
$$X_3 + X_5 + X_4 + X_7 = = 2543$$

Differenz der Summengleichung :

$$(X_3 + X_4 + X_5 + X_7) + X_6 = 2658$$
$$-(X_3 + X_4 + X_5 + X_7) = -2543$$
$$X_6 = 115$$

Ein hinsichtlich des Rechenzeitaufwandes mit dem Differenzenverfahren vergleichbares Prüfungsverfahren, das aber hinreichend für die Sicherung geheimer Werte in einer Untertabelle ist, leitet sich aus obigem Quaderkonzept ab. Danach werden nur solche geheimen Werte als gesichert angesehen, die einem n-dimensionalen Quader mit lauter geheimen Werten angehören.

Obiges Beispiels weist auch bereits auf die Lückenhaftigkeit von Abgleichsverfahren hin, wo einander überlappende Tabellen in einem iterativen Prozess solange aneinander abgeglichen werden, bis alle Tabellen in sich

sicher sind und dabei alle mehreren Tabellen gemeinsamen Werte in allen Tabellen, denen sie angehören, den selben Geheimhaltungsstatus haben:

Durch Zerlegung der Tabelle von Abb. 1.9 in alle eindimensionalen Tabellen, in ihre 5 Zeilen und ihre 5 Spalten, und Sicherung aller dieser eindimensionalen Tabellen mit iterativem Abgleich liefert genau obiges Sperrmuster. Dieses Sperrmuster ist eben nicht sicher, obwohl jede eindimensionale Tabelle für sich betrachtet (durch eindimensionale Quader) hinreichend geschützt ist und alle mehreren eindimensionalen Tabellen gemeinsamen Werte in jeder dieser Tabellen den selben Geheimhaltungsstatus haben.

Ähnliche Geheimhaltungslücken zeigen sich auch beim Untertabellenabgleichsverfahren, das bisher zur Überführung von hierarchisch gegliederten Statistiktabellen in zwischensummenfreie Tabellen angewendet wird. – Wie bemerkt, ist diese Umstrukturierung der gegebenen Tabelle in zwischensummenfreie Grundvoraussetzung für die Anwendbarkeit des Quaderverfahrens - . Eine eingehende Analyse des Untertabellenabgleichs und der dadurch verursachten Geheimhaltungslücken, findet man im Abschnitt 6.2. Zur Herleitung des Quaderverfahrens werden im Folgenden zunächst nur zwischensummenfreie Tabellen behandelt.

2. Vermeidung eindeutiger Rückrechenbarkeit

2.1 Allgemeine Einführung des Quaderkonzepts

Einen hinreichenden Schutz gegen eindeutige Rückrechnung geheimer Werte bietet ein Quaderverfahren, das die Prüf- und die Sperrfunktion in einem vereinigt: Es überprüft jeden primär geheimen Wert einer n-dimensionalen Untertabelle, ob er einem n-dimensionalen Quader mit lauter gesperrten Werten angehört (Prüffunktion des Quaderverfahrens), und es sichert ihn gegebenenfalls durch Sperren noch offener Quaderwerte (Sperrfunktion des Quaderverfahrens). Während das oben als Prüfverfahren vorgestellte Differenzenverfahren nur dann eine erforderliche Sicherung z.B. durch einen Quader zu sperrender Werte anzeigt, wenn der zu sichernde Wert zu einer Summe mit sonst lauter offenen Tabellenwerten als Summanden beiträgt, tritt beim Quaderverfahren mit Prüffunktion der Sicherungsfall bereits dann ein, wenn sich für das zu sichernde Tabellenfeld kein Quader mit lauter geheimen Quaderwerten finden lässt. Mit dem Begriff „Quaderverfahren“ ist fortan immer die o.g. Doppelfunktion angesprochen. Außerdem werden, wenn nicht anders erwähnt, Tabellen betrachtet, die nicht durch Zwischensummen unterteilt sind; anderenfalls könnte man das Geheimhaltungsproblem durch geeignete Umstrukturierung der Tabellendaten z.B. in eine Untertabellenhierarchie oder in vollständige Tabellen in zwischensummenfreie Tabellen überführen. Tabellen ohne Zwischensummen werden synonym als Untertabellen bezeichnet.

Das Quaderverfahren ist von besonderer praktischer Bedeutung,

- weil es bei nicht zu großen Tabellen (z.B. 10 000 Felder, 30 Untertabellen) sowohl maschinell als auch manuell durchgeführt werden kann; es besteht somit direkte manuelle Überprüfbarkeit;

- weil es auch n-dimensionale Tabellen von der Größenordnung 1 000 000 Tabellenfelder mit geringem Rechenzeitaufwand (im Bereich von CPU-Minuten) gegen Rückrechnung geheimer Werte sichern kann (ohne Dimensionsaufstockung gemäß 6.2.2) und

- weil es ein in dem Sinne optimales Verfahren ist, das für nur einen zu sichernden Wert die kleinste Anzahl von Partnerwerten auswählt, die diesen Wert in einer n-dimensionalen Untertabelle vollständig gegen Rückrechnung sichern.

Das Argument der manuellen Durchführbarkeit des Quaderverfahrens, erster Spiegelstrich, sollte keines Falls unterschätzt werden, bietet es doch die Möglichkeit einer direkten manuellen Überprüfung – zumindest für nicht zu große Teile von Tabellen – und schafft damit ein gewisses Vertrauen zum Ergebnis der Quadersicherung. Im Gegensatz dazu stelle man sich eine mit Hilfe eines sehr komplexen Algorithmus gesicherte Tabelle vor: Das Ergebnis so eines Sperrverfahrens kann der Nutzer im Allgemeinen nur zur Kenntnis nehmen; d.h. er wäre auch gar nicht verwundert, wenn ein ganz anderes Sperrmuster angezeigt worden wäre! ...

Neben dieser akzeptanzfördernden Wirkung der manuellen Durchführbarkeit des Quaderverfahrens ist auch die innovatorische von größter Bedeutung: Durch direktes Nachvollziehen oder auch durch rein manuelles Setzen von Sekundärsperrungen kann der Nutzer erfahren, zu welchen Problemen zu feine Gliederungen führen, wie die von

ihm eingetragenen Gewichtsfunktionen (vergleiche 5.3 im 2. Teil) das Sperrmuster verändern oder welchen Einfluss Vorinformationen auf die Sicherung von sensiblen Daten haben. Gerade von dieser innovatorischen Eigenschaft wird in der vorliegenden Arbeit ausgiebig Gebrauch gemacht, indem an Hand von vielen Beispielen die Weiterführung des Quaderverfahrens zu immer höherer Sicherheit der sensiblen Daten aufgezeigt wird.

Die im zweiten Spiegelstrich angesprochene hohe Bearbeitungsgeschwindigkeit hat ihre Ursache in einer für das Quaderverfahren ganz spezifischen Minimaleigenschaft, zum Schutze jedes einzelnen zu sichernden Wertes (Pivots) nur die kleinste Anzahl von Partnerwerten auszuwählen, die für einen hinreichenden Schutz des betreffenden Wertes ausreicht (dritter Spiegelstrich). Von allen Verfahren, die Tabellen auf der Basis von Einzelwertsicherungen gegen Rückrechnung ihrer geheimen Werte schützen, ist also das Quaderverfahren schon allein vom Prinzip her das schnellste, eben weil man zum Schutze jedes Pivots nur die kleinstmögliche Anzahl von Partnerwerten aufzusuchen braucht, die diesen Wert bereits hinreichend schützen.

Die im dritten Spiegelstrich aufgezeigte Minimaleigenschaft des Quaderverfahrens lässt sich folgendermaßen begründen: Ausgehend von einer eindimensionalen Tabelle mit einer wenigstens zwei geheime Werte umfassenden Schutzgesamtheit, in der das Pivot gesichert ist, wird sich die Anzahl der Partnerwerte mit jeder weiteren Gliederung mindestens verdoppeln, weil alle vormals geschützten geheimen Werte nach Erhöhung der Tabellendimension um eine Gliederung mindestens je einen Partnerwert pro vormals geheimen Wert zur Vermeidung der Rückrechnung bezüglich der neuen Gliederung benötigen. In einer n-dimensionalen Tabelle gehört ein gesicherter Pivotwert also einer Schutzgesamtheit von mindestens 2^n geheimen Werten an; ein Quader in einer n-dimensionalen Tabelle umfasst aber nur 2^n Tabellenwerte (vgl. 2.2.1.1), also die kleinste Anzahl von Partnerwerten, die für einen hinreichenden Schutz eines Pivots ausreichen.

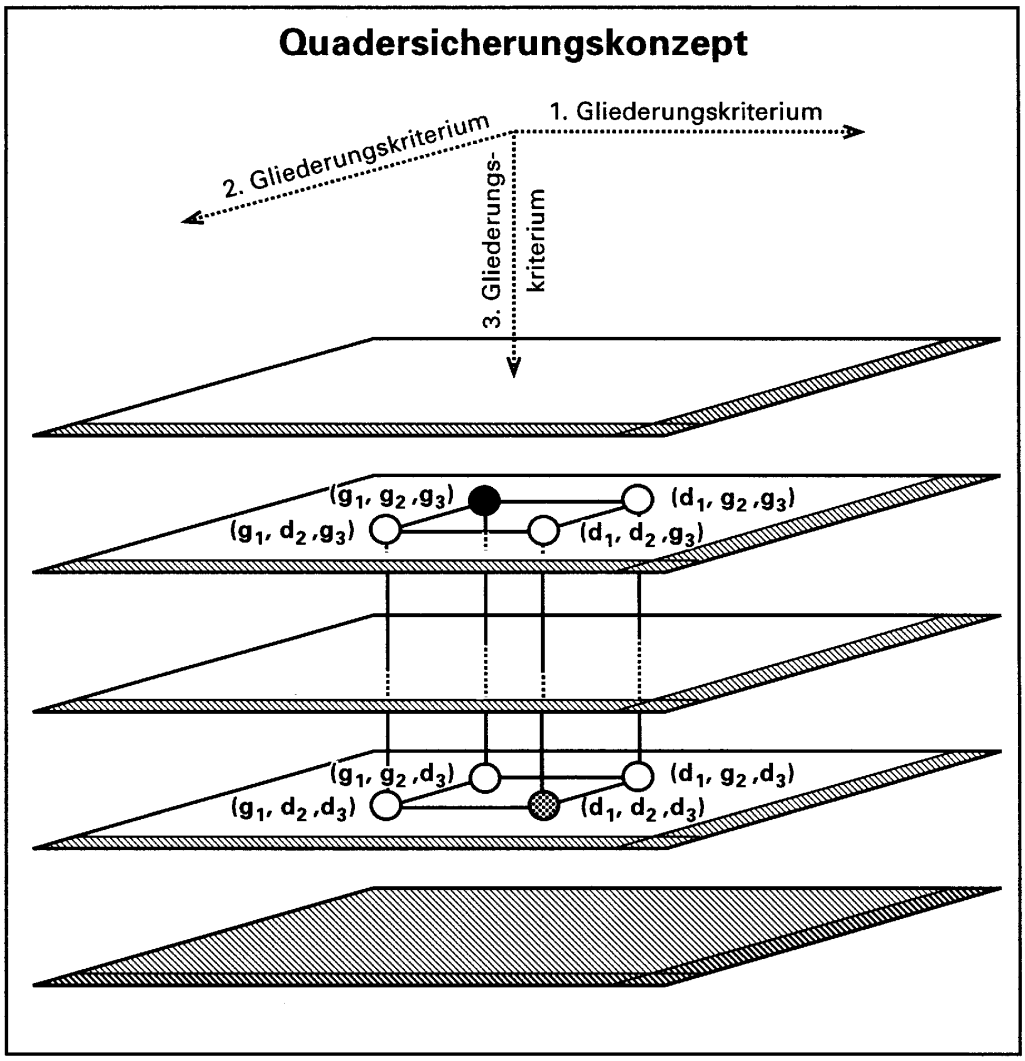
Bei allen Diskussionen von Sperrmustern ist zu bedenken, dass die Anzahl gesperrter Werte und damit auch die Anzahl der zu ihrer Berechnung aufzustellenden linearen Gleichungen bei realen Tabellen oft weit in die Hunderttausende geht. Um dennoch die Übersicht zu behalten, kann es sehr hilfreich sein, sich die Verhältnisse an Hand eines n-dimensionalen diskreten Raumes zu verdeutlichen; diese Sichtweise wird im Folgenden wesentlich unterstützt.

2.1.1 Allgemeine Definitionen und Regelungen

In diesem Abschnitt werden Null-Werte noch nicht als Quaderwerte zur Sicherung eines geheimen Wertes zugelassen, weil die Null als reproduzierendes Element in der Addition nur unter sehr einschränkenden Voraussetzungen als Sicherungspartner zu gebrauchen ist und hier zunächst die einfachste Situation behandelt werden soll. Auf die Probleme bei der Nutzung von Nullen wird in Abschnitt 3 und unter 5.1.2.3 ausführlich eingegangen.

Zur Erweiterung des im ersten Abschnitt anhand 2-dimensionaler Tabellen vorbereiteten Quadersicherungskonzepts auf n-dimensionale Tabellen betrachte man zunächst die in Abb. 2.1 schematisch dargestellte dreidimensionale Tabelle. Darin ist ein durch die Ausprägungen seiner drei Gliederungskriterien (Indizes (g_1, g_2, g_3)) fixierter geheimer Tabellenwert durch weitere geheime Werte (Primär- oder Sekundärsperrungen) in den Ecken eines dreidimensionalen Quaders gesichert worden.

Abb. 2.1



Um den geheimen Wert zuerst in seiner Ebene mit festgehaltenem Gliederungswert g_3 zu schützen, wird das Karree $K(g_3) = \{(g_1, g_2, g_3), (d_1, g_2, g_3), (g_1, d_2, g_3), (d_1, d_2, g_3)\}$ ausgewählt. Da eine Rückrechnung geheimer Werte auch über die dritte Gliederung erfolgen kann, wird noch ein weiteres Karree $K(d_3)$ als Projektion von $K(g_3)$ in die durch d_3 indizierte Ebene zur Sicherung der geheimen Werte von $K(g_3)$ aufgesucht, anschließend werden alle so festgelegten noch offenen Werte gesperrt.

Bei Betrachtung der 8 Indextripel des 3-dimensionalen Quaders sieht man, dass diese 2^3 Quadereckwerte durch das Tripel des geheimen zu sichernden Wertes (g_1, g_2, g_3) und das Tripel des dazu diametralen Wertes (d_1, d_2, d_3) eindeutig festgelegt sind, denn jeder Indexwert eines Quadereckwerttripels ist entweder der des zu sichernden oder der des dazu diametralen Wertes. Dabei werden alle $2*2*2=8$ Kombinationen durchlaufen.

Zur Übertragung dieses Quadersicherungskonzepts auf beliebige n-dimensionale Tabellen bedarf es folgender

Definitionen:

1. **Ein durch n Gliederungskriterien indizierter Tabellenwert heißt zu einem anderen diametral, wenn sich die Indizes beider Tabellenwerte in jedem Gliederungskriterium voneinander unterscheiden.**

2. **Die Gesamtheit aller durch n Gliederungskriterien indizierter Tabellenwerte, die durch zwei zueinander diametrale Werte so festgelegt ist, dass jeder Indexwert gleich dem entsprechenden Index eines der beiden Diametralwerte ist, heißt n-dimensionaler Quader.**

3. **Ein durch n Gliederungskriterien indizierter geheimer Wert (Einzelangabe oder auch nicht) heißt quadergesichert, wenn er zur Gesamtheit eines n-dimensionalen Quaders mit lauter von Null verschiedenen gesperrten Werten gehört, die – mit Ausnahme des zu schützenden Wertes selbst - keine Einzelangaben sind.**

Um die durch obige Definitionen fixierte Quadersicherung auf alle Tabellen anwenden zu können und sie außerdem noch im Sinne der Überlegungen von Abschnitt 1.1 zu optimieren, werden folgende Regeln angegeben:

1. Optimierungs-Regel:

Von allen Quadern, die mit dem zu sichernden Wert gebildet werden können, soll derjenige mit den meisten bereits gesperrten Werten ausgewählt werden. Wenn dann noch mehrere Quader zur Auswahl stehen, ist derjenige zu bevorzugen, der die kleinste Summe noch zu sperrender Werte aufweist.

2. Doppelquadersicherungs-Regel:

Enthält der auszuwählende Quader zur Sicherung eines geheimen Wertes außer dem zu schützenden Wert selbst mindestens eine Einzelangabe, so muss noch ein zweiter Quader zum Schutze des betreffenden Wertes aufgebaut werden, der jeden Einzelmelder im ersten Quader - mit Ausnahme im zu schützenden Werte selbst – ausschließt (Doppelquadersicherung).

3. Einzelangaben-Randsummen-Regel:

Eine Einzelangabe im Summenrand einer Tabelle ist wie eine primär geheime Angabe mit mehr als einem Merkmalsträger zu behandeln, die selbst nicht mehr gesichert werden muss (sie wird bereits durch einen Quader der zugehörigen Einzelangabe im Tabelleninneren geschützt).

Wenn nicht anders erwähnt, werden im nachfolgenden Text nur Tabellen behandelt, deren Werte von lauter verschiedenen Meldern stammen, so dass Einzelangaben in unterschiedlichen Tabellenfeldern auch unterschiedlichen Meldern angehören. Die zweite und dritte dieser Regeln werden noch im nachfolgenden Abschnitt begründet und eingehend erläutert.

2.1.2 Behandlung von Einzelangaben

2.1.2.1 Doppelquadersicherung

Nach Definition 3 sollten Einzelangaben im Allgemeinen keine "Sicherungspartner" in einem n-dimensionalen Quader sein, weil - wie noch gezeigt wird - die allgemeine Lösung der Quadergleichungen eine einparametrige Gesamtheit ist und somit jeder Merkmalsträger einer Einzelangabe seine Quaderwerte eindeutig berechnen kann, wenn sie nicht durch weitere Quader gesichert sind. Gleichwohl werden auch Einzelangaben durch Quadersperrungen gesichert, weil zwar der hier auftretende Merkmalsträger der Einzelangabe die nur zu seinem Schutz gesperrten Partnerwerte berechnen, umgekehrt aber kein anderer diese Werte aus den Quadergleichungen festlegen kann. Definition 3 ist aber für den Umgang mit Einzelangaben zu stringent.

Betrachtet man beispielsweise die zweidimensionale (Unter-)Tabelle der Abbildung 2.2, deren Werte mit Ausnahme der Randsummenwerte alle jeweils nur einem Merkmalsträger zugeordnet sind, so ist der allein durch den Quader {(1,A), (1,B), (2,A), (2,B)} geschützte Betrag 10 im Feld (1,A) nicht wirklich gesichert, weil jeder Merkmalsträger der anderen Quaderwerte aus der Kenntnis seines eigenen Wertes den Betrag 10 durch Differenzbildung mit dem zugehörigen Summenwert und den anderen hier wegen Einzelquadersicherung zunächst als offen angenommenen nicht zum obigen Quader gehörigen Werte berechnen könnte. Aber schon durch die Auswahl eines zweiten Quaders zum Schutze von (1,A), der die Einzelfälle des ersten Quaders - mit Ausnahme des zu schützenden Feldes - nicht enthält, etwa des Quaders {(1,A), (1,C), (3,A), (3,C)}, wird die eindeutige Rückrechenbarkeit des Wertes 10 von (1,A) verhindert: Jeder Merkmalsträger des einen Quaders könnte zwar aus dem Wissen seines Wertes alle anderen Werte seines Quaders berechnen, allein der jeweils andere Quader, von dem ihm kein Wert bekannt ist, hindert ihn daran.

Durch die Doppelquadersicherungs-Regel von 2.1.1 wird verhindert, dass durch eine rigorose (Einzel-) Quadersicherung gemäß Definition 3 in Tabellen der Gestalt der Abb. 2.2 alle Summenwerte gesperrt werden müssen! Diese Beispieltabelle ist bereits durch ihre Primärsperrungen gegen eindeutige Rückrechnung hinreichend gesichert.

Abb. 2.2 Einzelangaben sichern sich gegenseitig durch Doppelquader

	A	B	C	Σ
1	● 1 10	● 1 20	● 1 40	3 70
2	● 1 20	● 1 30	● 1 10	3 60
3	● 1 30	● 1 10	● 1 50	3 90
Σ	3 60	3 60	3 100	9 220

obere Zeile: Fallzahl
untere Zeile: Betrag

● = geheim zu haltender Wert

Die Doppelquadersicherung zur Vermeidung von Randsummensperrungen ist schon dann angezeigt, wenn bei nur einer Summierung von Tabellenwert-Aggregaten außer Einzelangaben keine Werte zu mehr als einem Berichtenden

auftreten. In obiger Tabelle ist beispielsweise der Wert 10 im Feld (1, A) mit der Doppelquadersicherung zu schützen, wenn ausschließlich die erste Zeile oder ausschließlich die erste Spalte lauter Einzelangaben im Tabelleninneren enthält, während die anderen Tabellenwerte von mehr als einem Berichtenden stammen. Dabei könnte der zu sichernde Wert selbst (hier 10) auch eine Primärsperrung zu mehr als einem Berichtenden sein.

2.1.2.2 Einzelangabe im Rand

Eine außergewöhnliche Situation bei der Behandlung von Einzelangaben liegt vor, wenn Randsummen selbst Einzelangaben sind. In so einem Fall ist dieselbe Einzelangabe sowohl im Rand als auch im Inneren der Tabelle anzutreffen; beide Werte werden lediglich durch ihre Indizes voneinander unterschieden. Bei der Sicherung einer dieser Einzelangaben lässt sich kein Quader finden, der nicht immer auch die entsprechende andere Einzelangabe enthielte! Fasst man also solche Einzelangaben aufgrund ihrer unterschiedlichen Indizierung als zwei verschiedene Tabellenwerte auf, so sind die vorgestellten Regeln des vorhergehenden Abschnitts 2.1.1 für den Schutz solcher geheimen Werte gar nicht anwendbar.

Das wird durch das Beispiel der Abbildung 2.3 besonders deutlich: Für die Sicherung der Einzelangabe im Inneren kommt überhaupt nur die entsprechende Einzelangabe in der Zeilen-, in der Spalten- und in der Eckfeldsumme in Frage. Hier ist offensichtlich eine Sicherung des Pivots ohne die Identifikation der vier Einzelmelder mit einem einzigen und der darauf basierenden Einzelangaben-Randsummen-Regel von 2.1.1 nicht zu machen.

Abb.: 2.3 Bei Fallzahladdition sind Einzelangaben im Rand mit ihrer "Quelle" im Tabelleninneren als identisch (paarig) anzunehmen

	A	B	C	Σ
1				
2	● 1 20			●1 20
3				
Σ	● 1 20			● 1 20

obere Zeile: Fallzahl
untere Zeile: Betrag

● = geheim zu haltender Wert

Sieht man aber in der im Rand und im Inneren stehenden Angabe ein und desselben Meldenden auch dieselbe Einzelangabe, so ist jeder Quader, der keine weiteren Einzelangaben zu anderen Meldern enthält als Sicherungsquader der betrachteten Einzelangabe geeignet; es genügt dann bereits eine einfache Einzelquadersicherung. Das liegt wiederum darin begründet, dass zwar der Einzelmerkmalsträger aufgrund der Kenntnis seines Wertes alle anderen Quaderwerte berechnen, umgekehrt aber keiner der anderen Merkmalsträger die Einzelangabe ermitteln kann - wie bereits oben festgestellt. Bei dieser Betrachtung ist es im Übrigen unerheblich, ob die betreffende Einzelangabe nur in einem oder wie in Abbildung 2.3 gleich in mehreren Rändern auftritt, weil sie ja immer als nur ein Tabellenwert in die Betrachtung einbezogen wird.

Demnach kann man jede Einzelangabe im Rand zusammen mit ihrer Quelle, der zugehörigen Einzelangabe im Tabelleninneren, mit einem einzigen Quader, dessen Pivot die Einzelangabe im Tabelleninneren ist, sichern. In diesem Sicherungsquader sind die Einzelangaben im Rand also wie gewöhnliche primär geheime Werte aufzufassen, die selbst nicht mehr gesichert werden müssen. Das genau besagt die dritte Regel von Abschnitt 2.1.1.

Diese Regel ist zulässig, weil die zu befürchtende Situation, dass eine Einzelangabe bei der Quadersicherung als Sicherungspartner benutzt wird, ohne sie als solche zu identifizieren, nicht eintreten kann: Gehört eine Einzelangabe im Tabellenrand einem Sicherungsquader an, so auch immer die zugehörige Einzelangabe im Inneren der Tabelle, weil in Bezug auf den betreffenden Summationsindex keine weitere Angabe im Tabelleninneren zu finden ist (Paarigkeit der Einzelangaben) - anderenfalls wäre der Randsummenwert keine Einzelangabe -. Das verhindert den Einsatz von Randeinzelangaben als Sicherungspartner bei der Einzelquadersicherung, es sei denn zum Eigenschutz desselben nur durch die Indizierung unterschiedenen Wertes.

Paarige Einzelangaben findet man immer, wenn bei der Summierung der Tabellenwerte gleichzeitig auch ihre Fallzahlen addiert werden und das war bisher stets vorausgesetzt worden. Es gibt aber auch Gliederungen in Statistiktabellen, bei denen die Werte allein, nicht aber die zugehörigen Fallzahlen addiert werden, wie beispielsweise die Gliederung des Umsatzes ein und desselben Unternehmens nach Inlandumsatz, Auslandumsatz und Gesamtumsatz. In solchen Gliederungen können mehrere Einzelangaben zu ein und der selben Einzelangabe aufsummiert werden. Die Paarigkeit solcher Einzelangaben mit denen im Rand ist dann nicht mehr garantiert. Bei Anwendung der Regel 3 kann die Identifizierbarkeit einer Randsumme als Einzelangabe u.U. gestört sein.

Um dennoch die Einzelangaben-Randsummen-Regel beizubehalten – sie ist bei Tabellen vom Typ der Abbildung 2.3 unverzichtbar - , ist sicherzustellen, dass bei nichtadditiven Fallzahlen eine Einzelangabe im Rand trotz Regel 3 noch erkannt wird. Dazu kann man solche Tabellen durch Hinzufügen der Randsummentabelle ohne Fallzahlsummierung zu einem „Pool“ aus der Tabelle selbst und evtl. weiteren Randsummentabellen ohne Fallzahlsummierung als sogenannte überlappende Tabellen behandeln (vgl.6.1). Siehe dazu Abbildung 2.4:

Abb. 2.4 “Tabellen-Pool” aus Originaltabelle und der ohne Fallzahladdition gebildeten Zeilensummentabelle

	A	B	C	Σ
1	● 1 10	● 2 20		⊙ 3 30
2	● 1 20	● 2 30		⊙ 3 50
3	● 1 30	● 2 10		⊙ 3 40
Σ	● 1 60	● 2 60		3 □ 120

obere Zeile: Fallzahl
untere Zeile: Betrag

● = primär geheimer Wert

⊙ = sekundär geheimer Wert

□ = Sekundärsperrung, erst nach Abgleich im Pool

Σ	● 1 60	● 2 60		⊙ 3 120

Weil Einzelangaben im Rand nach Regel 3, Abschnitt 2.1.1, als nicht zu sichernde Primärsperrungen gewertet werden, scheint die obere Tabelle der Abbildung 2.4 bereits durch ihre Primärsperrungen ● und die Sekundärsperrungen ⊙ im rechten Rand gesichert zu sein. Das trifft aber für die Summenzeile ohne den Sperreintrag □ nicht zu, weil der Wert in Spalte B durch die geheime Einzelangabe in Spalte A nicht geschützt ist. Erst in der abgetrennten Summenzeile als Einzeltabelle wird die Einzelangabe – jetzt im Tabelleninneren gelegen – als Einzelangabe erkannt und die Primärsperrung in Spalte B folgerichtig durch die Summensperrung ⊙ gesichert. Der Tabellenabgleich im Pool erzwingt dann mit □ die vollständige Sicherung auch der oberen Tabelle.

Allein auf die Tabellenwerte bezogene Summierungen findet man bei Tabellen, deren Felder mehr als einen Wert enthalten, wobei jedes Tabellenfeld auch noch die Summe seiner Werte ausweist. Nullen verhindern dann oft die einfache Sicherung eines „führenden Merkmals“ mit anschließendem „Durchstechen“ (alle Werte des selben Feldes erhalten den Geheimhaltungsstatus des führenden Merkmals), weil dabei einem von Null verschiedenen „führenden“ geheimen Wert u.U. eine nicht schützende geheime Null gegenübersteht; was bleibt, ist dann die „Pool-Behandlung“ nach obigem Muster.

Eine „Pool-Behandlung“ ist immer dann verzichtbar, wenn Einzelangaben in allen Summen mit ihren Angaben im Inneren paarig auftreten. Anders als in Abbildung 2.4 bereitet die Sicherung des Wertes 60 in der Abbildung 2.5 keine Offenlegungsprobleme: Ein Sicherungsquader zum Pivot (Σ, B) über das Feld (Σ, A), {(3, A), (3, B), (Σ, A), (Σ, B)}, hilft hier nicht weiter. Die Einzelangabe im Feld (Σ, A) wird zwar – weil im Rand gelegen - als solche nicht erkannt, aber die dazu paarige Angabe in (3, A) im Tabellen-Inneren wird sehr wohl als Einzelangabe gesehen. Wegen der Paarigkeit dieser Einzelangaben ist kein weiterer Sicherungspartner mit Elementen in Spalte A verfügbar. Daher ist über die Spalte A - anders als in Abbildung 2.4 - keine aufgrund der Einzelangabeneigenschaft von (3, A) erforderliche Doppelquadersicherung möglich; es muss über die Eckfeldsumme gesichert werden. Der Abgleich mit der Zeilensummentabelle wie in Abb. 2.4 ist daher nicht erforderlich.

Abb. 2.5 Paarigkeit von Einzelangaben im Rand und im Inneren erübrigt eine Pool-Behandlung

	A	B	C	Σ
1		● 2 20		● 2 20
2		● 2 30		● 2 30
3	● 1 30	● 2 10		⊙ 3 40
Σ	● 1 30	● 2 60		⊙ 3 90

obere Zeile: Fallzahl
untere Zeile: Betrag

● = primär geheimer Wert

⊙ = sekundär geheimer Wert

Während in der Beispieltabelle, Abb. 2.5, die für die Sekundärsperrung so wichtige Paarigkeit von Einzelangaben im Rand durch Nullwerte für die anderen beiden Gliederungswerte in Spalte A entsteht, können bei additiven Fallzahlen in allen Gliederungen niemals zwei oder mehr Einzelangaben zu einer Einzelangaben-Summe beitragen. In Tabellen mit ausschließlich additiven Fallzahlen kann es daher nur paarige Einzelangaben im Rand geben, so dass

das Problem der Randeinzelangaben in solchen Tabellen von vorneherein, d.h. ohne Pool-Behandlung durch Regel 3 von Abschnitt 2.1.1 vollständig gelöst ist.

Es ist hier noch darauf hinzuweisen, dass bei Fallzahl-Additivität in allen Gliederungen Regel 3 auch dann ihre Gültigkeit behält, wenn sogenannte Nullwerte als Sicherungspartner in einem n-dimensionalen Quader zugelassen sind (vgl. 3.1.2) und diese von einer von Null verschiedenen Anzahl von Berichtenden gemeldet wurden. Lediglich leere Tabellenfelder als Sperrkandidaten könnten Probleme bereiten. Damit könnte das paarweise Auftreten von Einzelangaben im Rand und im Inneren der Tabelle unkenntlich gemacht werden, weil anstelle der Einzelangabe im Tabelleninneren auch noch gewisse leere Felder als Sperrkandidaten zur Auswahl stünden wie z.B. das Feld (1, A) in Abbildung 2.5. Um dies zu vermeiden, kann man den sperrbaren leeren Feldern temporär von Null verschiedene Fallzahlen zuordnen, die bei der Fallzahladdition zu berücksichtigen sind. Dann haben sie für die Geheimhaltung die selbe Bedeutung wie die o.g. von tatsächlich vorhandenen Berichtenden gemeldeten Nullwerte.

Die Einzelangaben-Randsummen-Regel behält auch beim „Quaderverfahren in fiktiver vollständiger Tabelle“ (Abschnitt 7.), wo höher aggregierte Einzelangaben als Pivot in einem Strang geheimer Werte gesichert werden (vgl. 7.2) ihre Gültigkeit, wenn die anderen im Rand aber nicht in dem gerade bearbeiteten Strang des höher aggregierten Pivots liegenden Einzelangaben nach der Regel 3 von Abschnitt 2.1.1 behandelt werden. In obiger Beispieltabelle, Abb. 2.3, ist ein zum Pivot des Summeneckfeldes gehöriger Strang durch den Wert des Summeneckfeldes, den Summenwert der Spalte A und die Einzelangabe im Tabelleninneren als Diametralwert gegeben. Die einzige, nicht in diesem Strang, sondern im Summenrand der Zeile mit Gliederungsmerkmal 2 liegende Einzelangabe, wird dann wie eine „gewöhnliche“ Primärsperrung behandelt, die selbst nicht gesichert werden muss. Daher ist auch bei dieser Betrachtungsweise keine Doppelquadersicherung erforderlich.

2.2 Grundeigenschaften n-dimensionaler Quader

2.2.1 Die Quader-Index-Gesamtheit

Ein heuristisches Verfahren wie das Quaderverfahren erhält seine methodische Rechtfertigung gegenüber entsprechenden exakten Verfahren allein aus seiner praxisrelevanten technischen Realisierbarkeit. Methodik und Technik sind hier untrennbar verbunden und durchdringen einander. Das bedeutet u.a., dass methodische und organisatorische Überlegungen einander bedingen. – Die schnelle Auffindbarkeit eines Sicherungsquaders, eine der technischen Grundvoraussetzungen für einen praktikablen Einsatz des Quaderverfahrens, erzwingt den Aufbau eines effektiven Quaderindex-Algorithmus – Aspekt der technischen Realisierbarkeit – und motiviert gleichzeitig eine für das gesamte Quaderverfahren äußerst wichtige Aufteilung des Quaders in eine gerade und eine ungerade indizierte Teilgesamtheit (vgl. 3.1.2) – methodischer Aspekt.

2.2.1.1 Mächtigkeit eines n-dimensionalen Quaders

Das Index-n-Tupel g eines zu sichernden geheimen Wertes G sei durch

$$g = (g_1 , g_2 , ---, g_i , ---, g_n)$$

gegeben. Ein dazu diametraler Tabellenwert D habe die Indizes

$$d = (d_1 , d_2 , ---, d_i , ---, d_n)$$

mit

$$d_i \neq g_i \quad \text{für} \quad i = 1, 2, 3, ---, n$$

Die Ungleichung sichert, dass die beiden Tabellenwerte D und G zueinander diametral sind; d. h. dass die zum selben Gliederungskriterium i gehörigen Indizes d_i und g_i voneinander verschieden sind (Definition 1) und zwar für alle n Gliederungskriterien i = 1, 2, 3, ---, n.

Der zu den beiden zueinander diametralen Werten D und G gehörige Quader ist die Gesamtheit aller Tabellenwerte Q, die durch $\{q_i\}$ mit

$$q = (q_1 , q_2 , q_3 , ---, q_i , ---, q_n)$$

indiziert sind, wobei gilt (Definition 2):

$$q_i = \begin{cases} \text{entweder } g_i \\ \\ \text{oder } d_i \end{cases} \quad , i = 1, 2, 3, - - - , n.$$

Da dem gemäß jeder Indexwert q_i zum Gliederungskriterium i - unabhängig von den n-1 anderen - zwei Werte annehmen kann, den i-ten Indexwert g_i des zu sichernden geheimen oder den i-ten Index d_i des dazu diametralen Wertes, besteht der Quader aus 2^n Tabellenwerten.

2.2.1.2 Herleitung der Quader-Index-Formel

Um alle 2^n Quaderwerte aufsuchen zu können, ohne dabei n ineinander geschachtelte Schleifen durchlaufen zu müssen, wird der jeweils zu bearbeitende Quader auf einen Normquader abgebildet:

Der Normquader ist eine fiktive Gesamtheit n-fach indizierter Tabellenwerte, die durch die zueinander diametralen Werte mit n Nullen bzw. n Einsen als Index-n-Tupel definiert ist. - Bei dieser Hilfskonstruktion sind die Quaderwerte selbst ohne Belang; es kommt nur auf die Indizes an -. Diese Normquader-Index-Gesamtheit lässt sich demnach beschreiben durch

$$\{(B_1(k), B_2(k), B_3(k), ---, B_i(k), ---, B_n(k)), \quad k = 0, 1, 2, ---, 2^n - 1\},$$

wobei mit

$$B_i(k) = \begin{cases} \text{entweder } 0 \\ \\ \text{oder } 1 \end{cases} \quad \text{, für } i = 1, 2, 3, ---, n$$

eine binäre Variable eingeführt wurde und k die Nummer des betrachteten Normquaderwertes bezeichnet - in einer nun herzuleitenden Nummerierung -. Jeder Normquaderwert ist somit durch ein n-Tupel von Nullen und Einsen indiziert, die als Binärstellen der Nummer des betreffenden Wertes aufgefasst werden können. Das n-Tupel lässt sich damit in eine natürliche Dezimalzahl k umkodieren.

Ist z.B. (0,1,0,0,1) das Index-5-Tupel eines Normquaderwertes einer 5-dimensionalen Tabelle, so ist die Nummer dieses Normquaderwertes, wenn 01001 als binäre Darstellung der Nummer k aufgefasst wird, $k = 01001_{bin} = 0*2^4 + 1*2^3 + 0*2^2 + 0*2^1 + 1*2^0 = 9_{dez}$. Dem gemäß haben die den Normquader fixierenden zueinander diametralen Werte mit Indizes (0,0,0,0,0) und (1,1,1,1,1) die Normquaderwerte-Nummern k = 0 und k = 31.

Man erhält auf diese Weise eine ganz bestimmte mit Null beginnende Nummerierung aller Quaderwerte und zwar so, dass die zum Gliederungskriterium i gehörige Binärstelle des k-ten Quader-Wertes gerade $B_i(k)$ ist. So können alle Normquaderwerte in einer Schleife mit nur einem Schleifenindex k aufgefunden, d. h. ihre jeweils n Indizes zusammengestellt werden.

Der Übergang zu einem durch die Indizes eines geheimen Wertes $\{g_i\}$ und eines dazu diametralen Wertes $\{d_i\}$, i = 1, 2, 3, ---, n fixierten Quaders geschieht dann, indem man zum Beispiel den Normquaderwert mit Nummer k = 0 und Indizes (0,0,0, ---, 0) mit dem geheimen Pivot-Wert (g_1, g_2, g_3, ---, g_n) identifiziert und den dazu diametralen Normquaderwert mit Nummer $k = 2^n-1$ entsprechend (1,1,1, ---, 1) mit dem zum Pivot diametralen Quaderwert mit Indizes (d_1 , d_2 , ---, d_n). Dies geschieht dadurch, dass jeder Index $q_i(k)$ des realen Quaderwertes mit der Nummer k zum i-ten Ordnungskriterium mit dem Index des Normquaderwerts $B_i(k)$ so verknüpft wird, dass $q_i(k) = g_i$ immer $B_i(k) = 0$ und $q_i(k) = d_i$ immer $B_i(k) = 1$ zugeordnet ist: Für i = 1,2,---n gilt

$$q_i(k) = g_i \leftrightarrow B_i(k) = 0 \quad \wedge \quad q_i(k) = d_i \leftrightarrow B_i(k) = 1.$$

Das Index-n-Tupel des k-ten realen Quaderwertes ist dann gemäß der **Quader-Index-Formel**

$$q_i(k) = g_i + B_i(k) * (d_i - g_i) \qquad (1)$$

für i = 1, 2, 3, ---, n und k = 0, 1, 2, ---, $2^n - 1$ zu berechnen, wobei $B_i(k)$ die i-te Binärstelle des Laufindex-Wertes k zum i-ten Gliederungskriterium bezeichnet. Die Anwendbarkeit dieser Quader-Index-Formel setzt voraus, dass die Ausprägungen der Gliederungsmerkmale ganzzahlig sind; anderenfalls kann es hilfreich sein, sich eine temporäre Umindizierung in ganze Zahlen vorzustellen.

2.2.2 Abschätzung des Rechenaufwands beim Quaderverfahren

2.2.2.1 Einzelquadersicherung

Für die qualitative Beurteilung des Rechenzeitaufwandes zur Sicherung der geheimen Werte in einer n-dimensionalen Untertabelle ist die Anzahl der elementaren Rechenoperationen R wie Aufsuchen, Vergleichen und Addition von Tabellenwerten maßgebend. R wird bestimmt durch die Anzahl zu sichernder geheimer Werte N_g, die Anzahl der zum jeweiligen zu sichernden Wert aufzusuchenden Quader N_q sowie durch die Anzahl der Quaderwerte eines n-dimensionalen Quaders.

Die Anzahl der für einen einzigen zu sichernden Tabellenwert aufzusuchenden Quader ist identisch mit der Anzahl aller zu diesem geheimen Wert diametralen Tabellenwerte, da jeder dieser Quader gemäß Definition 2 durch „seinen" Diametralwert fixiert ist. Wenn jedes Gliederungsmerkmal i (i=1,2,---,n) der n-dimensionalen Untertabelle m_i Ausprägungen hat (Randsummen eingeschlossen), so kommen nach Definition 1 davon m_i-1 als diametrale Indizes in Frage. Insgesamt stehen bei n Gliederungskriterien also

$N_q = (m_1 - 1) * (m_2 - 1) * - - - * (m_n - 1)$ Quader zur Sicherung eines geheimen Wertes zur Auswahl.

Bezeichnet a die mittlere Anzahl der elementaren Rechenoperationen pro Tabellenwert, so ergibt sich die Gesamtzahl der Rechenoperationen R zu

$$R = a * N_g * (m_1 - 1) * (m_2 - 1) * - - - * (m_n - 1) * 2^n$$

Wenn man berücksichtigt, dass $N_g \le m_1 * m_2 * - - - * m_n = T$ ist, wo T für die Gesamtzahl aller Werte der Untertabelle steht, ergibt sich als Abschätzung $R < a * T^2 * 2^n$. Demnach nimmt der Rechenzeitaufwand mit der Anzahl n der Gliederungsmerkmale exponentiell zu, wächst aber mit dem Tabellenumfang T nur quadratisch an. Das erklärt die im Vergleich zu linearen Optimierungsverfahren kleinen Rechenzeiten beim Quaderverfahren (siehe dazu auch die Beispiele in Anhang A).

2.2.2.2 Doppelquadersicherung

Diese Abschätzung des Rechenzeitaufwandes betrifft den denkbar einfachsten Fall der Quadersicherung, bei der jeder primär geheime Wert durch nur einen Quader ohne Einzelangaben gemäß Definition 3 gesichert werden kann. Dazu muss man unterstellen, dass für jeden primär geheimen Wert der Tabelle ein Sicherungsquader im Tabelleninneren tatsächlich existiert, der mit Ausnahme des zu schützenden Wertes keine Einzelangaben enthält. Es sei nochmals angemerkt, dass die Voraussetzung auch schon bei weniger exotischen Tabellen als Abb. 2.2 verletzt ist, wenn z.B. im Inneren einer zwischensummenfreien zweidimensionalen Tabelle nur in einer Zeile oder Spalte lauter Einzelangaben stehen.

Ist der Einzelquaderschutz gemäß Definition 3 mit Quadern, die ganz im Tabelleninneren liegen, nicht zu machen, so kommt die Doppelquadersicherung zum Einsatz. D.h. aus der Gesamtheit aller Quader, die den zu schützenden Wert als Pivot-Element gemeinsam haben, muss gemäß der 2. Regel von Abschnitt 2.1.1 ein Quaderpaar ausge-

wählt werden, das – mit Ausnahme des zu schützenden Wertes – sonst keine gemeinsamen Einzelangaben hat. Da dazu aber u.U. sogar alle Quaderpaare aufgebaut und abgeprüft werden müssen, hat man – zumindest bei der Untersuchung, ob der geheime Pivotwert bereits gesichert ist – u.U. alle $N_q * (N_q - 1) / 2$ Quaderpaare tatsächlich zu bearbeiten. Das gilt insbesondere immer dann, wenn das betreffende Pivot noch nicht gesichert ist, weil diese Aussage erst nach der Bearbeitung auch des letzten Quaderpaares gemacht werden kann. Bei der Doppelquadersicherung nimmt der Rechenaufwand demnach annähernd mit der dritten Potenz des Tabellenumfangs zu (vergleiche die vorhergehende Abschätzung des Rechenaufwandes), wächst also erwartungsgemäß erheblich schneller als bei Einzelquadersicherung.

Hier ist allerdings anzumerken, dass nur wenige Primärsperrungen für eine Doppelquaderbildung in Frage kommen. Die Beispieltabelle des Abschnitts 1.2.2 enthält in der Untertabelle unterster Aggregation des zweiten Zeilen- und des ersten Spaltenstreifens nur Einzelangaben als primär geheime Werte, ohne dadurch eine Doppelquadersicherung zu erzwingen. Nur ganz spezielle Verteilungen von Einzelangaben, die dabei außerdem noch in größerer Menge auftreten müssen, machen eine Doppelquadersicherung überhaupt erst erforderlich (einige Spezialfälle wurden oben genannt). D.h. bei der Abschätzung des Rechenzeitaufwandes mit Berücksichtigung der Doppelquadersicherung wird man ein Polynom dritten Grades im Tabellenumfang T anzusetzen haben, wobei der Term mit der dritten Potenz von T nur als Korrekturglied fungiert. Lediglich bei sehr umfangreichen, nicht durch Zwischensummen unterteilten Tabellen, wie sie bei den im zweiten Teil dieser Darstellung zu behandelnden vollständigen Tabellen auftreten, wird der kubische Term gegenüber dem quadratischen hervortreten, so dass der Tabellenumfang in diesen Fällen die Rechenzeit noch verstärkt beeinflusst.

Um den Rechenaufwand ganz generell zu verringern, wurde die Anzahl der Auswahlmöglichkeiten von Doppelquadern in allen bisher realisierten EDV-Programmen zum Quaderverfahren stark verkürzt: Die Doppelquaderauswahl ist als doppelte Einzelquadersicherung ausgeführt, d.h. das Programm sucht zunächst einen Einzelquader aus, der das Pivot sichert, und falls dieser Quader außer dem Pivot noch Einzelangaben enthält, wird ein zweiter Quader, der außer dem Pivot die Einzelangaben des ersten Quades nicht enthält, aufgesucht. Diese doppelte Einzelquaderauswahl, bei der also nicht zu jedem „ersten“ Quader alle möglichen „zweiten“ aufgestellt werden, führt lediglich zu einer Verdoppelung des Rechenaufwandes, so dass bei zunehmendem Tabellenumfang auch bei Doppelquadersicherung nur eine quadratische Zunahme der Rechenzeit hinzunehmen ist. Damit werden dann auch bei großen Tabellen wie der Umsatzsteuerstatistik NRW 1994 noch akzeptable Rechenzeiten erreicht (siehe Anhang A).

3. Zum Intervallschutz beim Quaderverfahren

Die meisten Statistiken, die den Einsatz von Sekundärsperrverfahren erforderlich machen, weisen ausschließlich nicht negative Tabellenwerte aus (so genannte positive Tabellen). Wenn dem Nutzer solcher Tabellen a priori bekannt ist, dass die vorliegenden Tabellenwerte nur positiv oder Null sein können, besitzt er eine Zusatzinformation, die das Geheimhaltungsproblem wesentlich verschärft: Die Vermeidung der eindeutigen Rückrechenbarkeit bietet keinen ausreichenden Schutz; insbesondere genügt es oft nicht, sehr große Werte durch Sperren vergleichsweise kleiner Werte zu schützen, weil die Summe aus kleinem und großem Wert einen u.U. inakzeptabel genauen Schätzwert des unbekannten großen Tabellenwertes darstellt (Dominanz). Diese Problematik wird für das Quaderverfahren im Folgenden eingehend diskutiert. Dabei werden ausschließlich Tabellen mit nicht negativen Werten behandelt.

Eine weitere Verschärfung des Geheimhaltungsproblems ergibt sich, wenn unterstellt werden muss, dass der Nutzer für jeden Tabellenwert ein Schätzintervall angeben kann, das den tatsächlichen Wert überdeckt und dessen Intervallgrenzen u.U. nur um 40 % bis 60 % vom jeweiligen Tabellenwert abweichen. Bei dieser Art der Zusatzinformation sind nicht nur positive Tabellen mit einem verschärften Intervallschutz zu sichern, sondern auch Tabellen mit positiven und negativen Werten. Zuerst soll jedoch der Intervallschutz für die etwas einfacher zu handhabenden positiven Tabellen aus dem Quaderkonzept hergeleitet werden.

3.1 Spannweite geheimer Werte in positiven Tabellen

3.1.1 Ansatz zur Spannweitenberechnung mit Hilfe linearer Optimierung

Mit Hilfe der gegebenen Untertabelle kann der externe Tabellen-Nutzer ein lineares Gleichungssystem[1]

$$C\ X = B$$

für die t unbekannten primär wie sekundär geheimen Tabellenwerte,

$$X^T = (X_1\ , X_2\ , X_3\ \ , ---, X_t)$$

aufstellen mit gegebener Koeffizienten-Matrix C mit nur Nullen bzw. positiven und auch negativen Einsen als Elementen und gegebenem Konstanten-Vektor B. Die negativen Einsen werden durch Sperrungen in die jeweiligen Randsummen eingetragen.

Wenn dieses Gleichungssystem - wegen sekundärer Geheimhaltung - nicht eindeutig lösbar ist, löst er die 2 t linearen Optimierungsaufgaben (k = 1, 2, 3, ---, t, t steht für total, weil damit primäre wie sekundäre Sperrungen erfasst werden) unter der Voraussetzung nicht negativer Werte:

Minimiere X_k	maximiere X_k
$C X = B$	$C X = B$
$X_i \geq 0$ für i = 1, 2, 3, ---, t	$X_i \geq 0$ für i = 1, 2, 3, ---, t.

Auf diese Weise können die möglichen Werte der X_k eingegrenzt werden. Mit Hilfe der Lösungen $maxX_k$ und $minX_k$ erhält der externe Tabellen-Nutzer für jeden geheimen Wert X_k eine Spannweite

$$range_k = maxX_k - minX_k \quad (k = 1, 2, 3, ---, t)$$

Diese Spannweite kann nur Bruchteile von Prozent eines geheimen Wertes X_k betragen; sein Schutz ist dann nicht mehr gewährleistet.

Andererseits, kann man diese Optimierungsverfahren, bei denen für jede Unbekannte X_k die untere und die obere Schutzintervallgrenze berechnet wird und zwar für alle möglichen Ergebnisse der anderen X_i, benutzen, um einen geeigneten Satz von Sekundärsperrpositionen für einen hinreichenden Schutz der Primärsperrungen zu finden. Dazu gibt man sich eine Gesamt-Zielfunktion und einen Satz von Sekundärsperrungen vor und führt jede der 2t Einzeloptimierungen durch. Lassen sich dabei alle Primärsperrungen nur bis auf hinreichend große Schutzintervalle eingrenzen, so ist der betreffende Satz von Sekundärsperrpositionen ein brauchbarer. Von allen brauchbaren Sätzen von Sekundärsperrpositionen wird derjenige ausgewählt, der die vorgegebene Gesamt-Zielfunktion optimiert.

Zwar braucht nicht jede Einzeloptimierung unabhängig von den vorhergehenden durchgerechnet werden - man kann ausnutzen, dass sich alle Einzeloptimierungen nur in der jeweiligen Einzel-Zielfunktion voneinander unterscheiden -, dennoch ist dieses Vorgehen für die in der Praxis vorliegenden großen Tabellen viel zu aufwendig. Das Hauptproblem dabei ist die Organisation und Bearbeitung der großen Menge von Sekundärpositionssätzen! Um dies zu mechanisieren, schreibt man das Problem folgendermaßen um:

Für jeden primär geheimen Wert stellt man zwei Teilgleichungssysteme nach obigem Muster auf, eines für die obere und eines für die untere Schutzintervallgrenze als Unbekannte. Dabei geht noch als Nebenbedingung ein, dass die obere Intervallgrenzvariable nicht kleiner als der betreffende Tabellenwert oder, bei Primärsperrungen, nicht kleiner

---

[1] Eine sehr detaillierte Darstellung der linearen Optimierung des Geheimhaltungsproblems findet man in der Arbeit von J.Geurts.

als die vorgegebene obere Schutzintervallgrenze ist; für die untere Intervallgrenzvariable gilt Entsprechendes mit „nicht größer“ anstelle von „nicht kleiner“.

Außerdem führt man noch für jeden Tabellenwert eine Indikatorvariable ein, die den Wert 1 hat, falls der Tabellenwert gesperrt werden soll und die sonst verschwindet. Als Zielfunktion verwendet man die mit der Information jedes Tabellenwertes gewichtete Summe aller Indikatorfunktionen als Maß für den Informationsverlust durch die gesperrten Werte. Als Information eines Tabellenwerts wird beispielsweise bei positiven Tabellen häufig der Tabellenwert selbst angesehen. Mit diesem Ansatz ergibt sich ein gemischt ganzzahliges lineares Optimierungsproblem, dessen Rechenzeiten etwa exponentiell mit dem Tabellenumfang zunehmen; er ist daher für praktische Anwendungen unbrauchbar. Eine zwar nur näherungsweise optimale, dafür aber praktikable Alternative ist das Quadeverfahren.

3.1.2 Abschätzung der Spannweite in positiven Tabellen

Spannweiten geheimer Werte, die höchstens so groß wie die mit linearer Optimierung berechneten sind und die bei hinreichender Größe somit auch einen hinreichenden Schutz bieten, können bei Sicherung mit n-dimensionalen Quadern auf besonders einfache Weise bestimmt werden.

3.1.2.1 Quader im Tabelleninneren

Betrachtet wird ein n-dimensionaler Quader ohne Randsummenwerte (d.h. im Inneren einer Untertabelle), der durch die Indizes des zu sichernden Wertes G,

$$g = (g_1 , g_2 , g_3 , ---, g_n)$$

und die Indizes des dazu diametralen Wertes D,

$$d = (d_1 , d_2 , d_3 , ---, d_n)$$

fixiert ist.

Definition:

Ein Quaderwert X heiße gerade indiziert, wenn die Anzahl seiner Indizes

$$q = (q_1 , q_2 , q_3 , ---, q_n),$$

die mit den entsprechenden Indizes von D übereinstimmen, gerade ist, anderenfalls heiße er ungerade indiziert. D. h. ein Quaderwert ist gerade indiziert, wenn die Summe der Binärstellenwerte seiner Quaderwertnummer k gerade ist (siehe dazu die Quader-Index-Formel).

Beispiel:

In dem durch die beiden zueinander diametralen Werte (g_1,g_2,g_3), (d_1,d_2,d_3) der Abb. 2.1 fixierten 3-dimensionalen Quader sind die Werte indiziert durch (g_1,g_2,g_3), (d_1,d_2,g_3) , (d_1,g_2,d_3) und (g_1,d_2,d_3), die den Normquaderindizes (0,0,0), (1,1,0), (1,0,1) und (0,1,1) entsprechen, gerade indiziert, weil sie eine gerade Anzahl von d's bzw. eine gerade Binärstellensumme aufweisen. Die durch (d_1,g_2,g_3), (g_1,d_2,g_3) (g_1,g_2,d_3), (d_1,d_2,d_3) entsprechend (1,0,0), (0,1,0), (0,0,1), (1,1,1) indizierten Werte sind ungerade indiziert.

Zur Aufstellung des linearen Gleichungssystems für die 2^n Quaderwerte X als Unbekannte hat man gemäß der Summationsvorschrift der Untertabelle für jedes Gliederungskriterium i über alle Indexausprägungen (ohne Randsummenindex) zu summieren, wobei jeweils die anderen, nicht durch die i-te Gliederung bestimmten Indizes Quaderwertindizes sind, die bei dieser Summenbildung unverändert bleiben. Weil jeder Quaderwertindex bezüglich eines Gliederungskriteriums nur zwei Werte annehmen kann, tragen immer auch nur zwei Quaderwerte X, X' zur jeweiligen Randsumme bei. Alle linearen Gleichungen des o.g. Quaders haben daher die Gestalt:

$$X + X' = \sum \qquad (2)$$

$\sum$ bezeichnet die Quaderwerte-Summe zum i-ten Gliederungskriterium und zu einem fest vorgegebenen, das i-te Gliederungskriterium nicht enthaltenden n-1-Tupel von Quaderwertindizes, gegeben als Randsumme abzüglich aller anderen Summanden, die nicht zum o. g. Quader gehören.

Für jeden der n Summationsindizes i und für alle der 2^{n-1} den jeweiligen Summationsindex i nicht enthaltenden n-1-Tupel von Quaderwertindizes $(q_1 , ---, q_{i-1} , q_{i+1} , ---, q_n)$ lässt sich genau eine Gleichung der Gestalt (2) aufstellen. Dem gemäß gibt es insgesamt $n*2^{n-1}$ Gleichungen (2). Bei Tabellen, die nach mehr als zwei Merkmalen gegliedert sind (n > 2), hat man mehr Gleichungen als Unbekannte. Davon sind aber, wie sich aus den nun folgenden Betrachtungen ergibt, nur 2^n-1 voneinander unabhängig. - Die für diese Darstellung der Quadergleichungen (2) gewählte verkürzte Schreibweise vermeidet eine hier unnötige Überfrachtung der Variablen-Symbole mit langen Indexleisten und gestaltet damit die Formelbilder übersichtlicher und einprägsamer.[2]

Beispiel:

---

[2] Die Quaderdefinition ergibt für jedes Gliederungskriterium i mit m_i Ausprägungen unmittelbar $Xq_1,---,q_{i-1},\mathbf{g}_i,q_{i+1},---,q_n + Xq_1,---,q_{i-1},\mathbf{d}_i,q_{i+1},---,q_n = Bq_1,---,q_{i-1},.,q_{i+1},---,q_n - \sum Aq_1,---,q_{i-1},j, q_{i+1},---,q_n$, wobei die indizierten Werte A, X alle zur selben Randsumme B beitragen, und die Summe $\sum$ über alle j, j=1,2,---,m_i-1, $j \neq g_i$, $j \neq d_i$ zu erstrecken ist.

Für obigen dreidimensionalen Quader (Abb. 2.1) ergibt sich beispielsweise durch Summenbildung über das erste Gliederungskriterium bei festem zweiten und dritten Index z. B. g_2, d_3 :

$X_{\mathbf{g_1}, g_2, d_3} + X_{\mathbf{d_1}, g_2, d_3}$ + Summe aller nicht zum Quader gehörigen inneren Tabellenwerte mit festen Indizes g_2, d_3 = Randsumme $\bullet, g_2, d_3$. Insgesamt kann man für den dreidimensionalen Quader $3*2^{3-1} = 12$ Gleichungen der Gestalt (2) formulieren, wovon aber nur $2^3 - 1 = 7$ voneinander unabhängig sind. – Den 12 Quadergleichungen (2) entsprechen genau die 12 Kanten des dreidimensionalen Quaders. -

Die voneinander unabhängigen Quadergleichungen (2) sind demnach genau die Bestimmungsgleichungen von 3.1.1, nachdem dort die anderen, nicht zu obigem Quader gehörenden geheimen Werte eliminiert worden sind, so dass jede Lösung von (2) immer auch Lösung der Gleichungen von 3.1.1 sein muss.

Ist nun in einer Quadergleichung (2) X gerade indiziert, so ist X' - in der selben Gleichung - ungerade indiziert, denn beide Werte unterscheiden sich nur in dem Summations-Index, d.h. X' hat einen diametralen Indexwert d_i im Summations-Index i mehr oder weniger als X.

Wird der gerade indizierte Wert X durch

$$\hat{X} = X + \varepsilon \geq 0 \qquad (3a)$$

geschätzt, so muss der Schätzer des ungerade indizierten Quaderwertes X'

$$\hat{X}' = X' - \varepsilon \geq 0 \qquad (3b)$$

sein, damit obige Quadergleichung auch für die Quaderwert-Schätzer richtig bleibt[3].

Diese beiden Beziehungen gelten für alle Werte ein und desselben Quaders mit demselben ε-Wert als Schätzfehler:

Ein beliebiger z.B. gerade indizierter Wert Z des betrachteten Quaders kann von X ausgehend "erreicht" werden, indem ein Index von X nach dem anderen in den entsprechenden Index von Z umgesetzt wird. Man erhält so aufeinander folgende Quaderwerte X,X',Y,Y',---,Z,Z', von denen je zwei benachbarte Werte immer durch eine Quadergleichung der Gestalt (2) miteinander verknüpft sind: $X+X' = \sum_{XX'}$, $X'+Y = \sum_{X'Y,}$ $Y+Y' = \sum_{YY'}$, ---, $Z+Z' = \sum_{ZZ'}$. Weil in dieser Folge von Quadergleichungen immer je zwei benachbarte Gleichungen einen Quaderwert als Summanden gemeinsam haben, gilt für die jeweiligen Schätzer nach jeder einzelnen Indexumbe-

---

[3] Die obere Beschränkung der Quaderwerte-Schätzer durch die Quadersumme $\Sigma = X + X'$ im Falle positiver Tabellen braucht in (3a,b) nicht angegeben zu werden, sie folgt aus der vorausgesetzten Positivität der Tabelle, ist also in (3a,b) bereits latent enthalten: $0 \leq X + \varepsilon \leq \Sigma \quad \Leftrightarrow \quad 0 \leq X + \varepsilon \quad \wedge \quad 0 \leq X' - \varepsilon$

setzung immer eine der beiden Gleichungen (3) mit demselben ε-Wert, und zwar immer mit dem Pluszeichen bei gerader und immer mit dem Minuszeichen bei ungerader Indizierung:

$$\hat{X} = X + \varepsilon, \hat{X}' = X' - \varepsilon,$$
$$\hat{X}' = X' - \varepsilon, \hat{Y} = Y + \varepsilon,$$
$$\hat{Y} = Y + \varepsilon, \hat{Y}' = Y' - \varepsilon,$$
$$---, ---,$$

so dass schließlich auch $\hat{Z} = Z + \varepsilon$ und $\hat{Z}' = Z' - \varepsilon$ richtig ist[4]. Obige Gleichungen gelten also für alle Quaderwerte X, X' mit demselben Schätzfehler ε.

Die gemeinsame Lösung (3) aller durch (2) dargestellten Quadergleichungen enthält also genau einen Parameter, den Schätzfehler ε. Bei frei wählbarem ε bedeutet demnach der in Abschnitt 2 definierte Quaderschutz (Pt. 3 der Definitionen) einen hinreichenden Schutz gegen Rückrechnung geheimer Werte. In welchen Grenzen ε frei gewählt werden kann, hängt von den die Quaderwerte eingrenzenden Voraussetzungen (z.B. Vorwissen) ab.

Wenn keine weiteren Voraussetzungen über die Tabellenwerte, wie z.B. Nichtnegativität der Werte, zu berücksichtigen sind, so sind alle geheimen Quaderwerte - unabhängig von den Werten selbst - bereits hinreichend geschützt. Denn bei Zulässigkeit auch negativer Werte unterliegen die gemäß (3) zu berechnenden Schätzwerte der Quaderwerte X keinen Beschränkungen, d.h. ε kann beliebige Werte annehmen und die Quaderwerte-Schätzer auch. In diesem Fall genügt allein der Schutz gegen eindeutige Rückrechnung geheimer Werte, der bereits mit dem unter Pt. 2 beschriebenen Quaderverfahren gewährleistet wird.

Werden aber, wie im Allgemeinen üblich, nicht negative Tabellenwerte unterstellt, so folgt aus der Forderung, dass auch die Schätzwerte der geheimen Quaderwerte (3) nicht negativ sein dürfen, dass positive ε-Werte höchstens so groß wie der kleinste ungerade indizierte Quaderwert min X' und negative ε-Werte betragsmäßig höchstens so groß wie der kleinste gerade indizierte Quaderwert min X sein können. Sind also negative Schätzwerte auszuschließen, so muss für $\varepsilon \geq 0$

$$\varepsilon \leq \min X'$$

und für $\varepsilon < 0$ muss

$$|\varepsilon| \leq \min X$$

sein. Das heißt, dass $-\min X \leq \varepsilon \leq \min X'$ gilt. Mit der Lösung der Quadergleichungen (3a,b) folgt daraus

---

[4] Auf einen vollständigen Induktionsbeweis wird hier zu Gunsten einer besseren Übersichtlichkeit verzichtet (vgl. auch vorangehende Fußnote).

$$\hat{X} \in [X - \min X, X + \min X'], \qquad \hat{X}' \in [X' - \min X', X' + \min X] \tag{4}$$

Einem n-dimensionalen Quader im Inneren einer Untertabelle sind gemäß obiger Ungleichungen zwei Fehlerschranken zugeordnet, sein kleinster gerade indizierter Wert und sein kleinster ungerade indizierter Wert. Die Spannweite der Quaderschätzwerte (range), d.h. die Intervalllänge der im Folgenden als Schutzintervalle bezeichneten Schätzwertbereiche (4), ist daher für alle Quaderwerte des betrachteten Quaders, für gerade indizierte wie für ungerade indizierte, die gleiche:

$$\text{range} = \min X' + \min X \tag{5}$$

Diese Spannweite stellt also eine Quadereigenschaft dar. Sie dient im Folgenden als Quaderauswahlkriterium.

3.1.2.2 Quader mit Randsummen

Die obige Abschätzung der Spannweite gilt, wie bemerkt, nur für Quader im Inneren einer n-dimensionalen Untertabelle. Sollen auch Randsummenwerte als Quaderwerte X bzw. X" fungieren, so findet man unter den Quadergleichungen (2) auch solche, bei denen die eine der beiden benachbarten Unbekannten X, X" auf der linken, die andere auf der rechten Seite des Gleichheitszeichens steht (das kann u.U. auch auf alle Quadergleichungen zutreffen – siehe anschließendes Beispiel).

$$X + \sum A = X'' \tag{2'}$$

$\sum$**A** bezeichnet die Summe der zu diesem Summationsindex i und zu dem n-1-Tupel der Quaderwertindizes ohne den i-ten Index gehörigen Tabellenwerte ohne die unbekannten Tabellenwerte X, X" (vergleiche Fußnote 1). Der Schätzfehler ε hat hier für beide benachbarten Quaderschätzer das gleiche Vorzeichen, obwohl X in Bezug auf seine bisher definierte Indizierung einer anderen Quaderteilgesamtheit angehört als X".

$$\hat{X} = X + \varepsilon \; ; \qquad \hat{X}'' = X'' + \varepsilon \tag{3'}$$

Um dennoch die für die Programmierung so handliche Aufteilung in gerade und ungerade indizierte Quaderteilgesamtheiten beizubehalten, wird die Geradzahligkeit der Indizierung nicht mehr allein an der Anzahl von Null verschiedener Binärstellen gemessen, sondern zu dieser noch die Aggregationsstufen als zusätzliche Indizes addiert. Da sich die Aggregationsstufen zweier benachbarter Quaderwerte X, X" in einer Gleichung (2') um genau eine Aggregationsstufe voneinander unterscheiden, wird der Unterschied ihrer Indizierung durch die Addition ihrer Aggregationsstufen genau kompensiert, so dass X und X" in Gleichung (2') der selben Quaderteilgesamtheit angehören, wie es die Schätzfehlervorzeichen in (3') verlangen.

Beispiel:

Gegeben sei eine zweidimensionale positive Tabelle mit nur einem von Null verschiedenen primär geheimen Wert als Pivot im Inneren der Tabelle. In dieser Tabelle enthalten auch die Randspalte und –zeile sowie das Summeneckfeld nur diesen einen primär geheimen Wert. Der Sicherungsquader umfasst demgemäß das Pivotelement im Inneren der Tabelle mit den Indizes (g_1; g_2) und den Aggregationsstufen (1;1), das wegen 0 + 0 + 1 + 1 = 2 gerade indiziert ist, die beiden Randsummenwerte mit den Indizes (g_1; d_2), (d_1; g_2) und den Aggregationsstufen (1; 2), (2; 1), die daher ebenfalls gerade indiziert sind (für das erste der beiden Felder gilt 0 + 1 + 1 + 2 = 4), und das Summenfeld (d_1; d_2) mit den Aggregationsstufen (2; 2), das ebenfalls gerade indiziert ist (1 + 1 + 2 + 2 = 6). Der kleinste gerade indizierte Wert ist demnach der (einzige) Tabellenwert selbst. Da in dem hier betrachteten Quader offensichtlich kein ungerade indizierter Quaderwert existiert, gibt es auch keinen kleinsten dieser Werte, so dass das Intervall der gerade indizierten (und daher auch aller) Schätzwerte in (4) keine obere Beschränkung hat; der Schätzwert des primär geheimen Wertes kann beliebig aus dem Intervall $[0; \infty)$ ausgewählt, die Spannweite als beliebig groß angenommen werden. Dieses Ergebnis überrascht nicht, weil in dieser Tabelle alle Werte geheim sind und somit keine lineare Gleichung mit auch nur einem als offen ausgewiesenen Tabellenwert existiert.

Definition

> **Ein Wert eines n-dimensionalen Quaders ist gerade indiziert, wenn die Summe aus den Binärstellenwerten seiner Quaderwert-Nummer k und seiner Aggregationsstufen gerade ist, anderenfalls ist er ungerade indiziert (Aggregationsstufen in Einserschritten durchnummeriert).**

Mit dieser Vereinbarung behalten die Schätzwertintervalle und die Spannweite der geheimen Quaderwerte auch für Quader mit Randsummen ihre Gültigkeit. Und wenn dieser range-Wert nicht größer als der mittels linearer Optimierung zu berechnende ist, hat man damit ein Quaderauswahlkriterium gefunden, das einen hinreichenden Intervallschutz bietet.

3.1.2.3 Quaderspannweite als Auswahlkriterium

Zu vorgegebenem Prozentwert q werden nur solche Quader zur Sicherung eines von Null verschiedenen geheimen Wertes X zugelassen, für die

$$100 * \text{range} / X > q \qquad (6)$$

gilt. Im Falle X = 0 muss die Spannweite des Sicherungsquaders größer als ein vorzugebender absoluter Wert sein, z.B. größer als der kleinste von Null verschiedene nicht primär geheime Tabellenwert.

Ist im Falle von Null verschiedener zu sichernder primär geheimer Werte ein Prozentwert q von beispielsweise q = 80 % vorgegeben, so lässt die Auswahlregel (6) nur solche Quader zur Sicherung eines primär geheimen Wertes X zu, deren Spannweite, bezogen auf den zu sichernden Wert X, größer als 80 % ausfällt. Es werden danach nur solche Quader zum Schutze von X ausgewählt, deren kleinster gerade indizierter und kleinster ungerade indizierter

Wert in ihrer Summe größer als 0,8*X sind. Danach kommen bei Tabellen mit nicht-negativen Werten nicht alle Quader mit von Null verschiedenen Werten in Betracht, sondern nur diejenigen, deren kleinste Werte beider Teilgesamtheiten mit dem zu sichernden Wert vergleichbar sind. Dabei werden in der Regel Quader ausgesucht, deren Werte oft größer als der zu sichernde Wert selbst sind. Die relative Spannweite dieser Werte kann dann kleiner sein als der vorgegebene Prozentwert q. Für diese Werte ist aber kein Intervallschutz erforderlich, es sei denn, sie wären selbst primär geheim; dann werden sie beim weiteren Überprüfen durch andere Quader geschützt, die die Bedingung (6) erfüllen.

Es bleibt noch zu zeigen, dass das Schutzintervall $[X_i - minX, X_i + minX']$ eines beliebigen, z. B. gerade indizierten quadergeschützten Wertes X_i innerhalb der betreffenden Untertabelle mit keinem Verfahren zur Lösung linearer Gleichungssysteme, wie unter 3.1.1 dargestellt, weiter eingeengt werden kann.

Dazu geht man von der existierenden Gesamtlösung für die r unbekannten geheimen Werte $X_1, X_2, ..., X_j, ..., X_r$ aus, bei der jedem X_j sein realer Untertabellenwert zugewiesen wird.

Außer dieser (selbstverständlichen) Gesamtlösung genügen aber auch alle Werte dem Untertabellengleichungssystem von 3.1.1, die aus obigen dadurch entstehen, dass man die dem Quader zur Sicherung von X_i angehörenden Werte durch die die Quadergleichungen (2) erfüllenden Schätzwerte (3) ersetzt, während alle nicht zum Schutzquader gehörenden geheimen Werte ihre Tabellenwerte beibehalten. Gemäß (4) sind das alle Gesamtlösungen, bei denen X_i bei gerader Indizierung im Intervall

$$X_i \in [X_i - minX, X_i + minX']$$

liegt; X_i besitzt also ein Schutzintervall mit Intervalllänge

$$X_i + minX' - (X_i - minX) = minX' + minX = range.$$

Da diese Lösungsmenge in jeder Gesamtlösungsmenge einer nach 3.1.1 durchgeführten linearen Optimierung enthalten sein muss, wird das Schutzintervall $[X_i - minX, X_i + minX']$ niemals eingeengt, und man hat mit range eine Schutzintervalllänge gefunden, die nicht größer als eine mit linearer Optimierung berechnete ist.(Bei ungerader Indizierung von X_i wird analog argumentiert.)

Dass alle Werte ein und desselben Quaders dieselbe Spannweite haben, liegt an der Einparametrigkeit der Lösungen der Quadergleichungen. Erst diese Einheitlichkeit der Spannweiten aller Werte eines Quaders macht die Größe „Spannweite“ zu einer Quadereigenschaft. Es bleibt nun noch zu prüfen, ob diese Quadereigenschaft von der Wahl des den Quader definierenden Paares diametraler Werte abhängt, d.h. ob diese Wahl einen Einfluss auf die Aufteilung des Quaders in zwei komplementäre Teilgesamtheiten gerade oder ungerade indizierter Werte hat.

Ganz ähnlich wie bei der Begründung eines einheitlichen Quaderschätzfehlers ε lässt sich zeigen, dass zwei Quaderwerte nur dann zur selben Quaderteilgesamtheit gehören, wenn die Anzahl der Indexumbesetzungen plus Aggregationswechsel beim Übergang von einem der Quaderwerte zum anderen in ihrer Summe gerade ist, und diese

Anzahl ist unabhängig von dem den Quader fixierenden Paar diametraler Werte. Jedem Quader ist daher genau eine Spannweite (range) zugeordnet.

Damit ist nun sichergestellt, dass jeder gemäß 2.1.1, Definition 3, quadergeschützte Wert aus noch offenen Tabellenwerten (der betreffenden Untertabelle) höchstens bis auf seine Spannweite genau berechnet werden kann. Die Quadereigenschaft „Spannweite“ ist daher als Quaderauswahlkriterium zu verwenden: Bei für den Schutz geheimer Werte zulässigen Quadern wird die Spannweite durch einen vorzugebenden Prozentwert q gemäß (6) bzw. durch einen Absolutwert im Falle primär geheimer Nullen nach unten beschränkt.

Darüber hinaus bieten die Formeln (5) und (6) die Möglichkeit, auch Quaderelemente mit Wert Null als Schutzpartner für geheime Werte einzusetzen, vorausgesetzt, es ist nicht bekannt, dass der betreffende Wert Null ist:

Wird der Sicherungsquader so ausgewählt, dass Nullwerte (bzw. leere Tabellenfelder) nur einer der beiden Quaderteilgesamtheiten angehören - und dies können bis zu 50 % aller Quaderwerte sein -, so ist die Spannweite von Null verschieden und der Quader bietet einen hinreichenden Schutz gegen eindeutige Rückrechenbarkeit seiner Werte und bei Erfüllung obiger Auswahlformel (6) auch einen hinreichenden Intervallschutz.

Um einen hinreichenden Intervallschutz zu garantieren, wobei nicht bekannte Nullwerte einbezogen werden können, muss der dritte Punkt der eingangs gegebenen Sicherheitsdefinition wie folgt umformuliert werden:

3. Ein durch n Gliederungskriterien indizierter geheimer Tabellenwert gilt als gesichert, wenn er zur Gesamtheit eines n-dimensionalen Quaders mit lauter gesperrten Werten gehört, dessen Spannweite größer als eine vorgegebene Schranke für diesen Wert ist.

Anmerkungen

1. Da das Ziel der sekundären Geheimhaltung darin besteht, nur die primär geheimen Werte gegen zu genaue Rückrechnung zu schützen, wird die Quaderauswahl so vorgenommen, dass die auf den jeweils zu schützenden primär geheimen Wert bezogene Spannweite des Quaders größer als die einer relativen Mindestspannweite entsprechende vorgegebene Schranke q ausfällt. Diese Beschränkung auf den Vergleich der relativen Spannweite des zu schützenden primär geheimen Wertes mit der vorgegebenen Schutzschranke q erweitert die Auswahlmöglichkeiten unter den vorhandenen Quadern der Untertabelle ganz wesentlich: Wären immer alle Werte des jeweiligen zur Auswahl stehenden Quaders gegen zu genaues Rückrechnen zu schützen, also auch seine sekundär geheimen Werte, so ließen sich im statistischen Mittel nur Quader mit größeren Spannweiten, als für den Schutz des primär geheimen Wertes notwendig, heranziehen, weil die zur Sicherung benötigten anderen Werte des Quaders auch größer als der zu schützende geheime Wert selbst sein können.

2. Wenn die sekundäre Geheimhaltung mit Untertabellenabgleich erfolgt, ist ein hinreichender Intervallschutz zu vorgegebener Mindestspannweite in denjenigen Untertabellen, die Randsummensperrungen erfordern, prinzipiell nicht mehr zu garantieren: Sperrungen in einer Untertabellenrandsumme werden beim Unterta-

bellenabgleich in einer Untertabelle höherer Hierarchie gesichert. Dabei werden die Randsummenwerte durch die mit dem jeweiligen Geheimhaltungsverfahren bestimmten Schutzintervallgrenzen eingeengt; solche Randsummenwerte können nicht mehr alle positiven reellen Zahlen annehmen, sondern liegen in den mit dem Geheimhaltungsverfahren zu berechnenden Intervallen. Diese Eingrenzung von Randsummen durch eine Sicherung in höherer Hierarchie müsste bereits bei der Sperrung von Randwerten zur Sicherung der gerade zu bearbeitenden Untertabelle berücksichtigt werden, was aber unmöglich ist, weil die Sicherung dieser Sekundärsperrungen im Rand erst nach Abarbeitung der Untertabelle erfolgen kann, so dass die Sicherungsintervalle zum Zeitpunkt der Untertabellensicherung noch gar nicht vorliegen und daher bei der Festlegung der Sperrungen in der gerade bearbeiteten Untertabelle auch nicht berücksichtigt werden können; der Intervallschutz ist dann nicht gesichert. Diese Begründung basiert nicht auf speziellen das jeweilige Geheimhaltungsverfahren betreffenden Voraussetzungen, sondern gilt ganz allgemein für alle Sperrverfahren, die einen Intervallschutz bieten – also auch für das Quaderverfahren. Beim Quaderverfahren kann man allerdings die „Störung“ des Intervallschutzes noch etwas eingrenzen: Davon betroffen sind nur diejenigen Primärsperrungen, deren Sicherungsquader Werte in unterschiedlichen Untertabellen haben, die anderen nicht!

Hiermit ergibt sich ein starkes Argument für das eingangs erwähnte Verfahren zur Überführung einer mehrfach durch Zwischensummen untergliederten Statistiktabelle in eine solche, die frei von Zwischensummen ist, indem die Tabellendimension durch Einführung neuer Gliederungskriterien so weit aufgestockt wird, bis in der aufgestockten Tabelle keine Zwischensummen mehr vorkommen. – Zur Vermeidung des Untertabellenabgleichs müsste diese Dimensionsaufstockung also korrekterweise immer vorgenommen werden (siehe dazu die Ausführungen unter 6.2). Erst in der von Zwischensummen freien (u.U. hochdimensionalen) Tabelle lässt sich mit dem Quaderverfahren ein hinreichender Intervallschutz erreichen.

Wie durch den Untertabellenabgleich eingetragene Vorabinformationen in Form von Schätzintervallen für die Tabellenwerte bei der Quaderauswahl zu behandeln wären, wenn sie zum Zeitpunkt der Bearbeitung der betreffenden Untertabelle bereits vorlägen, wird für das Quaderverfahren im nachfolgenden Abschnitt, der sich mit der allgemeinen Berücksichtigung von Schätzintervallen auseinandersetzt, eingehend erläutert. Da diese Angaben z.Z. der Bearbeitung aber nicht vorliegen, ließe sich das Verfahren des Untertabellenabgleichs mit Intervallschutz allenfalls als iteratives Vorgehen retten, das dem ursprünglichen Iterationsprozess des Untertabellenabgleichs ohne Intervallschutz noch zu überlagern wäre. Dass damit dann auch noch kein hinreichender Schutz zu gewährleisten ist, wird in Abschnitt 6.2.1 durch ein Gegenbeispiel belegt. Für eine hinreichende Sicherung einer mehrfach durch Zwischensummen untergliederten Tabelle gegen zu genaue Rückrechnung ihrer primär geheimen Werte bleibt nur noch die Aufstockung der Tabellendimension (Abschnitt 6.2.2).

3.1.3 Beispieltabelle mit Intervallschutz und Nullwerten als Sperrpartner

Die in nachfolgend dargestellter Beispieltabelle (Abbildung 3.0) eingetragenen Sperrungen wurden unter der Voraussetzung nicht negativer Tabellenwerte bei Intervallschutz mit 125 % relativer Mindestspannweite und unter Einbeziehung von Nullwerten als Sperrkandidaten erzielt. Außerdem blieben Schutzintervalle beim Untertabellenabgleich unberücksichtigt. Betrachtet man in der so gesicherten Tabelle wieder die durch die Zwischensummenspalten AA, AB, AC abgegrenzten Spaltenstreifen und vergleicht sie mit denen der Abbildung 1.7, so fällt auf, dass lediglich der linke Spaltenstreifen einen Unterschied in den Sperrmustern beider Tabellen aufweist, die beiden anderen Spaltenstreifen sind mit denen der Tabelle Abbildung 1.7 deckungsgleich - trotz 125 % Intervallschutz und Einbeziehung von Nullen in den Sperrprozess.

Die vollständige Übereinstimmung der Sperrmuster im mittleren und rechten Streifen in beiden Tabellen erklärt sich zum einen daraus, dass die Werte der Primärsperrungen im Vergleich zu den anderen Werten der jeweiligen Untertabelle verhältnismäßig klein sind und, da nicht auf Nullen zurückgegriffen wurde, immer eine Spannweite entsteht, die wesentlich größer als der zu sichernde geheime Wert selbst ist. Bei der Karreeauswahl spielt somit das Summenkriterium, wonach die zusätzlich zu sperrenden Werte in jedem Quader so klein wie möglich sein sollen, die Hauptrolle. Dass dabei nicht von der Möglichkeit der Nullensperrung Gebrauch gemacht wird, liegt daran, dass die Verteilung der Nullwerte in zu akzeptierenden Quadern durch die geforderte Spannweite von 125 % sehr genau festgelegt ist:

Ein Quader, der eine von Null verschiedene Spannweite ausweisen soll, kann - wie bereits bemerkt - Nullwerte immer nur in einer seiner beiden Teilgesamtheiten aufnehmen. Wenn darüber hinaus eine Spannweite größer als Eins gefordert wird, stehen nicht mehr beide Teilgesamtheiten zur Auswahl, sondern nur noch die das zu schützende geheime Feld enthaltende Teilgesamtheit[5] darf einen Nullwert enthalten, die andere, ungerade indizierte Teilgesamtheit muss nullwertefrei bleiben. Enthielte die ungerade indizierte Quaderteilgesamtheit einen Nullwert, wäre die range allein durch die den zu schützenden geheimen Wert enthaltende Teilgesamtheit bestimmt und zwar durch den kleinsten Wert dieser Gesamtheit. Der kleinste Wert der gerade indizierten Gesamtheit ist aber höchstens so groß wie der geheime zu schützende Wert selbst, so dass die Spannweite eines Quaders mit einer Null in seiner ungerade indizierten Teilgesamtheit immer höchstens so groß wie der zu schützende Wert ist, die relative Spannweite dieses Wertes also niemals größer als Eins ausfallen kann.

In dieser Betrachtung wurde kein Bezug auf die Tabellendimension genommen, so dass allgemein gilt: ist die geforderte relative Mindestspannweite bei der Auswahl von Sicherungsquadern größer als Eins, so dürfen Nullwerte ausschließlich in der gerade indizierten, den zu schützenden Tabellenwert enthaltenden Teilgesamtheit des Siche-

---

[5] Die den zu sichernden geheimen Wert enthaltende Teilgesamtheit sei hier als gerade indiziert angenommen; anderenfalls wird genauso argumentiert, wobei immer nur „gerade indiziert“ durch „ungerade indiziert“ und „ungerade indiziert“ durch „gerade indiziert“ zu ersetzen ist.

rungsquaders vorkommen; liegen hingegen die Nullwerte in der ungerade indizierten Quaderteilgesamtheit, so wird die Spannweite des Quaders höchstens so groß wie der zu sichernde geheime Wert selbst sein.

Die Quaderauswahl mit Range-Kriterium soll nun exemplarisch anhand der obersten Untertabelle niedrigster Aggregationsstufen im rechten Spaltenstreifen verdeutlicht werden. Die Verteilung der Sperrvermerke in den anderen Untertabellen des mittleren und des rechten Spaltenstreifens erklärt sich ganz ähnlich: Sie ist durch die Anordnung der Nullen in Verbindung mit dem Range-Kriterium bei relativer Mindestspannweite größer Eins, wonach Nullen nur in der den zu sichernden Wert enthaltenden Quaderteilgesamtheit auftreten dürfen, schon weitgehend fixiert.

Um als erstes das Range-Kriterium für den ausgewählten Sicherungsquader des primär geheimen Wertes 95 im Feld (134; AAA) zu verifizieren, ist zu beachten, dass die gerade indizierten Tabellenwerte beide oder keinen der Indizes des zu sichernden Wertes aufweisen - im ersten Fall hat der Wert keinen, im zweiten Fall zwei diametrale Indizes: Der zu sichernde Wert 95 ist demnach gerade indiziert (die Aggregationsstufensumme ist 1+1, also gerade für alle vier Quaderwerte), ebenso verhält es sich mit dem dazu diametralen Wert 34, der keinen Index mit dem zu sichernden Wert gemeinsam hat, also zwei diametrale Indizes besitzt. Die beiden anderen Karreewerte 321 und 256 sind demnach ungerade indiziert.

Man sieht, in einem Karree im Inneren einer Untertabelle gehören die auf einer Diagonale einander gegenüber liegenden Werte immer zur selben Quaderteilgesamtheit. Der kleinste gerade indizierte Wert des Sicherungsquaders von 95 ist daher 34, der kleinste ungerade indizierte Wert 256. Daraus erhält man die Quaderspannweite range = 34 + 256 = 290. Da die relative Spannweite des zu schützenden Wertes somit 290/95 = 3,05 beträgt, wird die relative Mindestspannweite von 1,25 überschritten, der ausgewählte Sicherungsquader ist in Bezug auf das Range-Kriterium zu akzeptieren. Es bleibt zu klären, ob noch ein anderer Quader mit kleinerer Summe zusätzlich zu sperrender Werte ebenfalls das Range-Kriterium erfüllt.

Dass der Versuch, Nullwerte in einen Sicherungsquader des primär geheimen Wertes 95 einzubeziehen, scheitern muss, liegt daran, dass die wegen des Range-Kriteriums nur in Betracht kommenden Karrees mit Nullwert als Diametralwert, das sind die Diametralfelder (131; AAC), (132; AAC), (131, AAB), (133; AAB), immer zu Karrees mit Nullen auch in der anderen ungerade indizierten Teilgesamtheit führen. Ihre Quaderspannweite ist daher immer Null. Die anderen noch als Sicherungsquader zu betrachtenden Karrees mit von Null verschiedenen Diametralwerten 732, 644, 432 oder auch 234 haben alle einen Nullwert in der ungerade indizierten Teilgesamtheit, nämlich in der Spalte AAA und führen somit zu einer relativen Spannweite, die nicht größer als Eins ist. Bei einer vorgegebenen Mindestspannweite von 1,25 sind auch diese Karrees inakzeptabel. Das einzige brauchbare Sicherungskarree ist das mit den Feldmarkierungen S und P.

Der linke Spaltenstreifen weist im Vergleich zur Beispieltabelle ohne range-Auswahl und ohne Einbeziehung von Nullwerten wesentlich weniger Sperrungen aus. Besonders bemerkenswert ist die Vermeidung von Sekundärsperrungen in den höher aggregierten Tabellenfeldern; dies ist auf die Möglichkeit, auch Nullwerte als Sperrpartner zu verwenden, zurückzuführen. Denn durch die Wahl des Nullwertes im Feld (112; ACC) als Diametralelement zum primär geheimen Wert 53, kann auf die Summensperrungen (110; ACB), (110; ACD) verzichtet werden. Der kleinste ungerade indizierte Wert in diesem Karree ist 221, so dass sich eine relative Spannweite für den zu si-

chernden geheimen Wert 53 von 221/53 = 4,17 ergibt, die größer als die vorgegebene von 1,25 ist. Die anderen fünf noch in Frage kommenden Karrees mit Diametralfeldern (112; ACB), (112; ACA), (111; ACC), (111; ACB), (111; ACA) haben als kleinsten ungerade indizierten Wert entweder immer 28, wenn der Diametralwert in der Zeile 111 liegt, oder 29 bzw. 423, wenn die Zeilennummer 112 Diametralindex ist.

Die Quader mit kleinstem ungerade indizierten Wert 28 bzw. 29 scheiden als Sicherungsquader aus, weil die diesem Wert entsprechenden ranges kleiner als der zu sichernde Wert 53 sind und damit die relative Mindestspannweite 1,25 unterschritten wird. Auch der Quader mit Diametralfeld (112; ACA) mit kleinstem ungerade indizierten Wert 423 scheidet als Sicherungsquader aus, weil seine Summe zusätzlich zu sperrender Werte 1001 + 423 = 1424 größer als die entsprechende Summe von 221 + 423 = 644 des ausgewählten mit S- und P-Markierungen versehenen Sicherungskarrees des primär geheimen Wertes 53 ist. Das Karree {(112; ACC), (112; ACD), (113; ACC), (113; ACD)} ist unter allen das günstigste; es wird allen Kriterien gerecht.

Sicherung der Beispieltabelle mit Intervallschutz und Nullwerten als Sperrpartner Abb. 3.0

1. Schlüssel	2. Schlüssel														
	ACD	**ACC**	**ACB**	**ACA**	**AC**	**ABC**	**ABB**	**ABA**	**AB**	**AAD**	**AAC**	**AAB**	**AAA**	**AA**	**A**
00000134	112 5 **S**	10 2 P	1.445 20	549 12 **S**	2.116 39	4.128 34	345 15	211 12	4.684 61	321 21 **S**	0 0	0 0	95 2 P	416 23	7.216 123
00000133	40 1 P	66 4 **S**	0 0	23 3 **S**	129 8	2.567 44	2.332 30	432 21	5.331 95	732 51	644 34	0 0	0 0	1.376 85	6.836 188
00000132	723 9	254 11	327 5	543 19	1.847 44	1.123 64	4.427 59	1.632 26	7.182 149	432 23	0 0	234 36	0 0	666 59	9.695 252
00000131	2.156 33	1.342 23	1.111 17	99 4	4.708 77	590 11	2.334 28	342 9	3.266 48	34 3 **S**	0 0	0 0	256 17 **S**	290 20	8.264 145
00000130	3.031 48	1.672 40	2.883 42	1.214 38	8.800 168	8.408 153	9.438 132	2.617 68	20.463 353	1.519 98	644 34	234 36	351 19	2.748 187	32.011 708
00000125	321 5 **S**	11 3 **S**	411 18	0 0	743 26	0 0	56 5	0 0	56 5	712 50	3.421 84	0 0	0 0	4.133 134	4.932 165
00000124	56 4 **S**	12 1 P	2.152 29	399 11	2.619 45	0 0	123 10	0 0	123 10	345 44	2.612 61	55 3	0 0	3.012 108	5.754 163
00000123	99 8	311 10	754 19	345 16	1.509 53	221 7	34 2 P	73 6 **S**	328 15	123 23	321 41	567 32	43 4	1.054 100	2.891 168
00000122	1.837 33 **S**	19 1 P	88 4	0 0	1.944 38	0 0	621 13	0 0	621 13	1.015 89	2.221 52	96 18	641 8	3.973 167	6.538 218
00000121	344 15	298 13	0 0	934 9	1.576 37	0 0	74 8	0 0	74 8	0 0	231 33	0 0	875 72	1.106 105	2.756 150
00000120	2.657 65	651 28	3.405 70	1.678 36	8.391 199	221 7	908 38 **S**	73 6 **S**	1.202 51	2.195 206	8.806 271	718 53	1.559 84	13.278 614	22.871 864
00000113	53 2 P	221 8 **S**	29 3	1.001 19	1.304 32	0 0	0 0	0 0	0 0	11 2 P	0 0	21 2 P	0 0	32 4	1.336 36
00000112	423 18 **S**	0 0 **S**	0 0	0 0	423 18	0 0	261 5 **S**	34 2 P	295 7	745 71	0 0	67 8	0 0	812 79	1.530 104
00000111	28 5	0 0	0 0	0 0	28 5	0 0	0 0	0 0	0 0	148 25 **S**	0 0	81 7 **S**	0 0	229 32	257 37
00000110	504 25	221 8	29 3	1.001 19	1.755 55	0 0	261 5 **S**	34 2 P	295 7	904 98	0 0	169 17	0 0	1.073 115	3.123 177
00000100	6.192 138	2.544 76	6.317 115	3.893 93	18.946 422	8.629 160	10.607 175	2.724 76	21.960 411	4.618 402	9.450 305	1.121 106	1.910 103	17.099 916	58.005 1.749

Legende: Wert / Berichtspfl. — 10.000 / 100 P — Sperrvermerk (P=primär, S=sekundär)

3.2 Berücksichtigung von Vorwissen

3.2.1 Externe Schätzintervalle bei der Spannweitenberechnung

3.2.1.1 Verschärfung des Geheimhaltungsproblems durch Vorwissen

Bei der Sicherung sensibler Daten in einer Veröffentlichungstabelle muss man bedenken, dass die Tabellennutzer selbst zum Kreis der Berichtenden und zum Kreis der zu Schützenden gehören. Sie verfügen somit in Bezug auf die Tabellendaten über besondere Kenntnisse und haben ein berechtigtes Interesse, dass ihre Anonymität trotz solchen Vorwissens durch die Tabellenveröffentlichung nicht aufgehoben wird. Der erweiterte Schutz von Einzelangaben sowie die Berücksichtigung der Positivität einer Tabelle in der bisher schon praktizierten sekundären Geheimhaltung sind Ausdruck der Anerkennung dieses Schutzbedürfnisses. Die bisherigen Sicherungsmaßnahmen bleiben aber höchst unbefriedigend, wenn man bedenkt, dass die Nutzer von Statistiktabellen aufgrund ihrer Erfahrung von Berufs wegen sehr viel mehr über „ihre“ Tabellen wissen als nur, dass sie keine negativen Werte enthalten. In der Regel kann man davon ausgehen, dass die Tabellenwerte zumindest bis auf plus minus 50 % genau bekannt sind.[6] Dann reicht aber der nur für positive Tabellenwerte hergeleitete Intervallschutz nicht mehr aus. Dies macht schon folgendes sehr einfache Beispiel deutlich:

Gegeben sei die zweidimensionale Tabelle der Abbildung 3.1 mit dem primär geheimen Wert 100 (Sperrvermerk p, Fallzahlen weggelassen)

Abb. 3.1

100 p	80	180
90	1	91
190	81	271

Werden in dieser Tabelle - mit irgendeinem Verfahren - die vier inneren Werte 100, 80, 90, 1 gesperrt, so sichert den primär geheimen Wert 100 gemäß (5) ein Schutzintervall mit range = 80 + 1 = 81 bzw. eine relative Spannweite von 81 %. Kann der Datennutzer aber seine Zusatzinformation in Form von Schätzintervallen einbringen, deren Intervallgrenzen um plus minus 50 % vom tatsächlichen Tabellenwert abweichen, so liegt ihm bei gesperrtem Tabelleninneren folgende „Intervalltabelle“ vor:

---

6 Kritik am Umgang der statistischen Ämter Kanadas, der USA und Europas mit sensiblen Daten von G. Sande. Gordon Sande (Fa. Sande & Associates, Inc.) hat für das statistische Amt Kanadas das Geheimhaltungsprogramm CONFID entwickelt und vertreibt eine von ihm weiterentwickelte Version auf kommerzieller Basis.

Abb. 3.2

[50; 150]	[40; 120]	180
[45;135]	[0,5;1,5]	91
190	81	271

Mit Hilfe dieser Schätzintervalle und den Summenbeziehungen der Tabelle findet er dann für den primär geheimen Wert das „Schutzintervall“

$$99{,}5 \leq X_1 \leq 100{,}5$$

oder eine relative Spannweite von 1 %. Der primär geheime Wert ist daher zu genau bestimmt (vergleiche auch das Beispiel von G. Sande, Fußnotenhinweis).

Zur Begründung des obigen 1 %-Intervalls kann man zunächst aus der Tabelle mit den eingetragenen Sperrvermerken X_1, X_2, X_3, und X_4

Abb. 3.3

X_1	X_2	180
X_3	X_4	91
190	81	271

die unbekannten X_2, X_3, X_4 mit Hilfe der Randsummenwerte eliminieren und erhält eine Tabelle mit einer einparametrigen Lösungsgesamtheit für die gesperrten Werte (siehe vorheriges Kapitel). Der zu schätzende Parameter ist hier X_1.

Abb. 3.4

X_1	$180-X_1$	180
$190-X_1$	$91-(190-X_1)$	91
190	81	271

Zur Eingrenzung von X_1 muss man diese Tabellenwerte mit den externen Schätzintervallrändern der Tabelle Abb. 3.2 vergleichen:

$50 \leq \quad X_1 \quad \leq 150$ $\qquad\qquad$ $50 \quad \leq X_1 \leq 150$

$40 \leq 180 - X_1 \leq 120$ $\qquad\qquad$ $60 \quad \leq X_1 \leq 140$

$45 \le 190 - X_1 \le 135$ oder $55 \le X_1 \le 145$
$0{,}5 \le 91-(190-X_1) \le 1{,}5$ $99{,}5 \le X_1 \le 100{,}5$

Der Nutzer wählt daraus diejenigen Intervallgrenzen von X_1 aus, die X_1 am genauesten festlegen, nämlich das letzte der vier Intervalle. Bei der Zusatzinformation, „es liegt eine positive Tabelle vor“, hätte er hingegen die Intervalle gebildet (weil er als externer Nutzer nicht auf die range-Formel (5) des Schutzquaders hätte zurückgreifen können)

Zeilen: $0 \le X_1 \le 180$ Spalten: $0 \le X_1 \le 190$
$0 \le 180 - X_1 \le 180$ $0 \le 180 - X_1 \le 81$
$0 \le 190 - X_1 \le 91$ $0 \le 190 - X_1 \le 190$
$0 \le 91-(190-X_1) \le 91$ $0 \le 91-(190-X_1) \le 81$

und er hätte dann daraus als kleinstes Intervall

$$99 \le X_1 \le 180$$

gefunden mit der relativen Spannweite range / X_1 = 81 %.

Man sieht: In diesem Beispiel genügt schon das Zusatzwissen, das die Tabellenwerte um lediglich plus minus 50 % eingrenzt, um damit den zu schützenden Wert 100 bis auf plus minus 0,5 % genau vorherzusagen.

3.2.1.2 Allgemeiner Ansatz zur Berücksichtigung von Vorwissen im Quaderverfahren

Die hier nur anhand eines speziellen Beispiels für den Intervallschutz bei Vorliegen von zusätzlicher Information vorgeführte Abschätzung der gesperrten Werte wird nun mit Hilfe des Quaderansatzes (also aus der Sicht des Tabellenschützers) auf beliebige n-dimensionale Tabellen verallgemeinert:

Als Vorinformation über aggregierte Tabellendaten wird im Folgenden das Wissen des Tabellennutzers bezeichnet, mit dem er in der Lage ist, ohne Kenntnis der Veröffentlichungstabelle Tabellenwerte abzuschätzen. Dem gemäß kann Vorinformation in Form von Schätzintervallen, die die tatsächlichen Tabellenwerte überdecken, dargestellt werden. Dazu gibt man zu jedem Tabellenwert X einen oberen X_o und einen unteren Schätzwert X_u an, der den vor Offenlegung der Tabelle noch unbekannten Tabellenwert eingrenzt, so dass für jeden Schätzwert $\hat{x}$ des tatsächlichen Wertes X gilt:

$$X_u \le \hat{x} \le X_o\,, \quad X_u = X_u(X)\,, \quad X_o = X_o(X)$$

Wird zur Sicherung der sensiblen Tabellenwerte das Quaderverfahren eingesetzt, so ist $\hat{x}$ auch als Lösung der Quadergleichungen zu behandeln: Q bezeichnet die durch Definition 2 im zweiten Kapitel gegebene Gesamtheit der Tabellenwerte eines n-dimensionalen Quaders, Q_g die Gesamtheit seiner gerade indizierten, Q_u die Gesamtheit seiner ungerade indizierten Werte (gemäß Definition des Abschnitts 3.1.2.2). Für alle X, $X' \in Q$ gelten also

die Gleichungen (3a) und (3b), wenn die beiden benachbarten Quaderwerte dieselbe Aggregationsstufen-Summe (Aggregationsstufen über alle n Gliederungen summiert) aufweisen bzw. (3´), wenn zwischen den beiden benachbarten Quaderwerten ein Wechsel in ihren Aggregationsstufen-Summen vorliegt. Die Schätzwerte $\hat{x}$ bzw. $\hat{x}'$ der gerade bzw. der ungerade indizierten Quaderwerte sind dem gemäß

$$\hat{x} = X + \varepsilon\,,\ \hat{x}' = X' - \varepsilon \quad ,$$

mit einem für alle Tabellenwerte des betrachteten Sicherungsquaders Q einheitlichen Schätzfehler ε und zwar für die gerade indizierten $X \in Q_g$ wie auch für die ungerade indizierten Quaderwerte $X' \in Q_u$. Der Schätzfehler ε, ist wieder der Parameter der einparametrigen Lösungsgesamtheit der Quadergleichungen; er wird durch das obige externe Schätzintervall eingeengt gemäß der Abschätzungen $X_u \le X + \varepsilon \le X_o$, $X'_u \le X' - \varepsilon \le X'_o$ für alle gerade und ungerade indizierten Quaderwerte, bzw.

$$-(X - X_u) \le \varepsilon \le X_o - X\,, \qquad -(X' - X'_u) \le -\varepsilon \le X'_o - X',$$

wobei man für die externen Schätzintervallgrenzen $X_{u,o} = X_{u,o}(X)$, $X \in Q_g$ bzw. $X'_{u,o} = X'_{u,o}(X')$, $X' \in Q_u$ zu berücksichtigen hat, d.h. dass diese Grenzen von den Werten selbst abhängig sind, die sie eingrenzen. Der „Quaderparameter“ ε ist also durch die Abstände der tatsächlichen Quaderwerte von ihren durch die Vorinformation bestimmten Intervallgrenzen beschränkt. Da ε für all diese Ungleichungen den gleichen Wert besitzt, ergibt sich
für positive ε-Werte gemäß $\varepsilon \le \min_{X \in Q_g} (X_o - X)$ und $\varepsilon \le \min_{X' \in Q_u} (X' - X'_u)$
ein oberer Schrankenwert $\varepsilon_+ \ge 0$

$$\varepsilon_+ = \min \left[\min_{X \in Q_g} (X_o - X), \min_{X' \in Q_u} (X' - X'_u) \right] \qquad (7a)$$

und für die absoluten Beträge der negativen ε-Werte $|\varepsilon| \le \min_{X' \in Q_u} (X'_o - X')$ und $|\varepsilon| \le \min_{X \in Q_g} (X - X_u)$

die obere Schranke $\varepsilon_- \ge 0$ (min strukturiert, kann aber durch nur ein „äußeres“ min ersetzt werden (Anhang C))

$$\varepsilon_- = \min \left[\min_{X' \in Q_u} (X'_o - X'), \min_{X \in Q_g} (X - X_u) \right] \qquad (7b)$$

Wenn eine der beiden Quaderteilgesamtheiten nicht existiert (Quader erstreckt sich über das Eckfeld wie z. B. in Abb. 2.3, Abschnitt 2.1.2.2), so ist das betreffende nicht angebbare Argument in der äußeren min-Funktion von (7a) oder (7b) fortzulassen also beispielsweise in (7a) nur $\min(X_o - X)$ oder $\min(X' - X'_u)$ zu verwenden.

Die vier minimalen Abstände, die kleinsten Abstände der tatsächlichen Tabellenwerte von ihren unteren und ihren oberen Intervallgrenzen der externen Vorinformation für die gerade und die ungerade indizierte Quaderteilgesamtheit, sind in aller Regel voneinander verschieden, so dass sich meist auch unterschiedliche Werte für die Schranken ε_+ und ε_- ergeben. Der Quaderparameter ε liegt demnach in dem asymmetrischen Intervall

$$-\varepsilon_{-} \leq \varepsilon \leq \varepsilon_{+}\,, \qquad \varepsilon_{-},\ \varepsilon_{+} \geq 0 \tag{8'}$$

mit der Spannweite

$$\text{range} = \varepsilon_{+} + \varepsilon_{-} \tag{8}$$

Diese Spannweite wurde ohne die Voraussetzung nicht negativer Tabellenwerte hergeleitet; sie gilt also auch für Tabellen, die sowohl positive als auch negative Werte enthalten können. Aus den Schrankengleichungen (7a), (7b) und der Beziehung für die Spannweite (8) der Quaderwerteschätzer folgen unmittelbar die entsprechenden ε-Schrankenwerte und die range des Quaderverfahrens für Tabellen mit der ausschließlichen Vorinformation nicht negativer Werte, wenn man die oberen Intervallgrenzen gegen unendlich gehen lässt[7] und die unteren Null setzt. **Das Quaderverfahren für sogenannte positive Tabellen stellt somit einen Spezialfall eines allgemeinen Quaderverfahrens dar, das eine allgemeine Vorinformation über die Tabellenwerte in Form von Schätzintervallen in den durch obige Schrankenwerte für ε beschriebenen Intervallschutz umsetzt.**

Im obigen Beispiel (Abb. 3.1) sind die Abstände der tatsächlichen Tabellenwerte von ihren Schätzintervallgrenzen im Falle der gerade indizierten Quaderteilgesamtheit 50;50 für den primär geheimen Wert und 0,5;0,5 für den dazu diametralen Wert. Für die beiden ungerade indizierten Werte hat man die Abweichungen 40;40 und 45;45. Die obere Fehlergrenze ε_{+} berechnet sich mit (7a) dem gemäß, $\varepsilon_{+} = \min [0,5;40] = 0,5$ und die untere Fehlergrenze ε_nach (7b) zu $\varepsilon_{\_} = \min [40;0,5] = 0,5$, in Übereinstimmung mit oben aufgeführten direkten Berechnungen.

Bei Berücksichtigung von externen Schätzfehlern bestimmt i. Allgm. der kleinste Quaderwert mit seinen im Vergleich zu den anderen Quaderwerten besonders kleinen Abweichungen von den Schätzintervallgrenzen die Quaderspannweite, während bei Berücksichtigung nur der Positivität der Tabelle sowohl der kleinste gerade indizierte als auch der kleinste ungerade indizierte Tabellenwert direkt in die Spannweitenberechnung eingehen. Bei der Minimierung der Abweichungen der tatsächlichen Tabellenwerte von ihren Schätzintervallgrenzen werden sowohl bei der oberen Fehlerschranke (ε_{+}) als auch bei der unteren (ε_) immer beide Quaderteilgesamtheiten durchlaufen und nicht wie bei der „nur positiven“ Tabelle bei ε_{+} nur die ungerade indizierten und bei ε_ nur die gerade indizierten Quaderwerte. (Eine Zusammenfassung des Verfahren-Kerns findet man im Anhang C.)

Es sei ausdrücklich angemerkt, dass die drastische Range-Einengung in diesem Beispiel nicht von der Eigenart symmetrischer Schätzintervalle herrührt. Die starke Verkürzung der range wird vielmehr durch die mit besonders kleinen Werten unweigerlich verbundenen sehr kleinen Schätzfehlerbeträge verursacht und zwar unabhängig davon,
ob das betreffende Schätzintervall symmetrisch ist oder nicht. Die Besonderheiten symmetrischer oder nahezu

---

[7] Die Tabellenpositivität beinhaltet bereits die obere Beschränkung der Quaderwerte (vgl. Fußnote zu (3a.b)) und braucht folglich hier nicht berücksichtigt zu werden: Ein externer Nutzer kann eine obere Quaderwertgrenze immer nur bis zur oberen Schutzintervallgrenze einengen (vgl. Herleitung der Quaderspannweite), d.h. in positiven Tabellen $X_o \geq X + \min X'$ und $X'_o \geq X' + \min X$. Eingesetzt in (7a,b) folgt mit (8) als untere Grenze für die Spannweite positiver Tabellen die Formel (5).

symmetrischer Schätzintervallgrenzen, für die obige Tabelle (Abb. 3.1 mit Abb. 3.2) ebenfalls ein Beispiel ist, werden im anschließenden Abschnitt näher erläutert.

3.2.2 Einbeziehung von Nullwerten bei symmetrischen Schätzintervallen

3.2.2.1 Symmetrische Schätzintervalle

Um den Intervallschutz mit dem Quaderverfahren bei Berücksichtigung von Vorinformationen auch EDV-mäßig zu gewährleisten, müssten zur Berechnung der Schrankenwerte außer dem Tabellenwert selbst noch zwei weitere Werte in jedes Tabellenfeld, d.h. in jeden Datensatz eingetragen werden. Dann ließe sich mit den Formeln für die untere und die obere ε-Schranke die range gemäß (8) auch in diesem allgemeinen Fall berechnen und damit eine geeignete Quaderauswahl treffen (ganz analog zum bisherigen Quaderverfahren für positive Tabellen). Unter diesen Umständen müsste man bei der Quaderauswahl auf drei verschiedene Einzeltabellen zugreifen, auf die Wertetabelle mit den bereits eingearbeiteten „Geheimhaltungsattributen“, auf die Tabelle der unteren und auf die der oberen Abweichung des tatsächlichen Tabellenwertes von der jeweiligen externen Schätzintervallgrenze oder auf die der Schätzintervallgrenzen selbst.

Wesentlich einfachere Verhältnisse liegen bei Tabellen mit zum Tabellenwert symmetrischen Schätzintervallen vor. Hier genügt die Angabe nur eines zusätzlichen Wertes, des Abweichungsbetrags des Tabellenwertes von einer der beiden Schätzintervallgrenzen. Dabei kann sogar die ursprüngliche Tabellenstruktur, bei der jedes Tabellenfeld durch die Ausprägungen seiner Gliederungsmerkmale, die Anzahl der Berichtenden, den Sperrschlüssel sowie durch den Tabellenwert charakterisiert ist, beibehalten werden, wenn man den Tabellenwert und seinen Abweichungsbetrag von der unteren oder oberen Schätzintervallgrenze als komplexe Zahl zusammenfasst. Der Tabellenwert wird z.B. dem Realteil, der Betrag seiner Abweichung von einer der beiden Schätzintervallgrenzen dem Imaginärteil der komplexen Zahl zugeordnet.

Leider sind die externen Schätzintervalle, die der Nutzer der Tabellendaten zur Eingrenzung der Werte angeben kann, in der Regel nicht symmetrisch angelegt, denn sonst könnte er einen sehr genauen Tabellenschätzwert angeben, den Mittelwert der betreffenden Schätzintervallgrenzen. Stattdessen kann man aber die ursprünglichen, vom Nutzer vorgegebenen Schätzintervalle durch kleinere, vom Nutzer-Schätzintervall überdeckte, zu den tatsächlichen Tabellenwerten symmetrische Intervalle approximieren und mit diesen die Quaderspannweite (8) berechnen. Wenn der Intervallschutz bei Verwendung der kleineren symmetrischen Approximationsintervalle gewährleistet werden kann, dann ist er erst recht bei den größeren Nutzerintervallen gegeben.

Um nun mit nur einer weiteren externen Angabe im Eingabestand auszukommen, kann man das neue Eingabefeld zur Eingabe des jeweils kleinsten Abweichungsbetrags des tatsächlichen Tabellenwertes (eingetragen im Wertefeld) von seinen Schätzintervallgrenzen der Vorinformation, ε_{min} (X), nutzen, also den Wert

$$\varepsilon_{min}(X) = \min [X_o - X, X - X_u]$$

in das neue Eingabefeld eintragen. Damit lassen sich die Schrankenwerte ε_+ und ε_- der obigen Schrankenformeln wie folgt nach unten abschätzen:

$$\varepsilon_+ \geq \min [\min_{X \in Q_g} \varepsilon_{min}(X), \min_{X' \in Q_u} \varepsilon_{min}(X')] \text{ und } \varepsilon_- \geq \min [\min_{X \in Q_g} \varepsilon_{min}(X), \min_{X' \in Q_u} \varepsilon_{min}(X')],$$

was wegen $Q = Q_g \cup Q_u$ zu einem für alle Tabellenwerte desselben Quaders Q einheitlichen Schrankenwert ε_Q führt, der die obere und die untere Fehlerschranke jedes Quaderwertes $X \in Q$ von unten beschränkt:

$$\varepsilon_Q = \min_{X \in Q} \varepsilon_{min}(X) \leq \min [\varepsilon_+, \varepsilon_-] \qquad (9)$$

Die damit zu berechnende Quaderspannweite eines Sicherungsquaders mit lauter gesperrten Werten

$$\text{range} = 2\varepsilon_Q$$

ist demnach kleiner als die zunächst für gegebene unsymmetrische externe Schätzintervalle hergeleitete. Die Schrankenwerte $\pm\varepsilon_Q$ grenzen den Parameter ε der Lösung der Quadergleichungen noch weiter ein als zuvor angegeben, so dass ein externer Tabellennutzer bei seiner Berechnung von Fehlerschranken die Quaderspannweite $2\varepsilon_Q$ auch unter Berücksichtigung von gegebener Vorinformationen nicht unterschreiten kann. Hat man also einen Quader zur Sicherung eines primär geheimen Wertes so ausgewählt, dass seine Spannweite $2\varepsilon_Q$ bezogen auf den primär geheimen Wert größer als eine vorgegebene Schranke ist (die man für den Schutz des primär geheimen Wertes für ausreichend hält), so ist nach Sperrung der noch offenen Quaderwerte ein hinreichender Intervallschutz garantiert.

Die Sicherung von primär geheimen Tabellenwerten bei Berücksichtigung von Vorinformationen mit Hilfe der zuletzt genannten Beziehungen kann jetzt auch optional mit dem EDV-Programm GHQUAR.4 durchgeführt werden, wenn man in ein zusätzlich im Datenbestand eingerichtetes Wertefeld den jeweils kleinsten absoluten Betrag des Schätzfehlers, $\varepsilon_{min}(X)$, einträgt. Dabei ist zu beachten, dass die benutzte Beziehung die Fehlerschranken des betreffenden Sicherungsquaders nur approximiert, was auch durch die Symmetrie des zuletzt angegebenen Schutzintervalls zum Ausdruck kommt. Die Erfüllung der Beziehung (6) mit $\text{range} = 2\varepsilon_Q$ beim Vergleich der relativen Spannweite mit einer vorgegebenen Schutzschranke muss in positiven Tabellen bei Quadern mit Nullen zwangsläufig zur Ablehnung solcher Quader zur Sicherung geheimer Werte führen, weil die Abweichung der Null von ihrem unteren externen Schätzintervallwert in positiven Tabellen immer Null ist und somit $\varepsilon_Q = 0$ sein muss.

3.2.2.2 Ergänzung symmetrischer Schätzintervalle für Nullwerte

Anders als die Formel (9) im vorangegangenen Abschnitt liefern die „exakten" Fehlerschrankenformeln (7a), (7b) auch unsymmetrische Quaderintervalle, $\varepsilon_+ \neq \varepsilon_-$, so dass die eine Schranke verschwinden, die andere aber dennoch von Null verschieden sein kann. Ein Nullwert mit von Null verschiedenem oberen Schätzfehler trägt nur in eine der beiden Fehlerschranken (7a), (7b) eine verschwindende Schätzfehlergrenze ein, die andere Fehlerschranke wird durch diesen Nullwert nicht zu Null und somit auch nicht die Quaderspannweite. **Nullwerte können also auch bei Vorliegen externer Vorinformation ganz legitime Schutzpartner primär geheimer Werte sein, wenn ihre obere Schätzfehlerschranke nur groß genug ist!**

Um bei der Sicherung geheimer Tabellenwerte mit nur einem Schätzfehler-Eingabewert pro Tabellenwert (symmetrische Schätzintervalle) auch Nullen mit einbeziehen zu können, ohne bei den Nullen die externe Vorinformation unberücksichtigt lassen zu müssen, ist die näherungsweise Berechnung der Spannweite so zu modifizieren, dass auch wieder unsymmetrische Quaderintervalle möglich werden. Es bietet sich an, die Schnittmengen der Schutzintervalle positiver Tabellen mit für Nullwerte modifizierten Näherungsintervallen zu verwenden und somit

$$\varepsilon_{Q+} = \min\,[\varepsilon_Q, \min_{X' \in Q_u} X'], \qquad \varepsilon_{Q-} = \min\,[\varepsilon_Q, \min_{X \in Q_g} X] \tag{10}$$

als Schrankenwerte einzuführen. Diese Schranken sind bei Quadern mit ausschließlich positiven Werten wegen $\min(X,X') \geq \varepsilon_Q$ keinesfalls größer als die einheitliche Schranke ε_Q (Hinzufügen von weiteren bei der Auswahl des kleinsten Wertes zusätzlich zu berücksichtigenden Argumenten in der min-Funktion führen höchstens zu kleineren Minimalwerten). Darüber hinaus muss dafür gesorgt werden, dass ε_Q nicht verschwindet, wenn Nullwerte in dem betreffenden Quader vorkommen. Dazu empfiehlt es sich, den oberen Intervallschätzer der externen Vorinformation, $X_0(0) > 0$, als „kleinsten" Schätzfehler des Nullwertes in das dafür vorgesehene Feld des Eingabedatenbestandes einzutragen, denn das ist genau diejenige Intervallgrenze der Null, die auch in den exakten Schrankenformeln (7a), (7b) wirksam ist. D.h. anstelle von (9) gilt nun

$$\varepsilon_Q = \min_{X \in Q \setminus 0} [\min \varepsilon_{min}(X), X_0(0)] \tag{9'}.$$

Mit dieser Näherung ist nun wieder ein asymmetrisches Quaderintervall eingeführt worden, das durch die exakten Schrankengleichungen nicht eingeengt wird und das somit als für den Intervallschutz hinreichend anzunehmen ist. Mit ihrer Hilfe lässt sich die Quaderspannweite in gewohnter Weise berechnen:

$$\text{range} = \varepsilon_{Q+} + \varepsilon_{Q-} \tag{11}$$

(11) besitzt die gewünschten Eigenschaften, dass beim Auftreten von Nullwerten die betroffene Fehlergrenze verschwindet. Enthält aber nur die eine der Quaderteilgesamtheiten Q_g, Q_u Nullwerte, so ist die Spannweite im All-

gemeinen von Null verschieden, der Quader bietet Intervallschutz – ganz analog zum bisherigen Quaderverfahren bei positiven Tabellen.[8]

3.2.3 Schätzintervalleintrag durch andere Tabellen oder Tabellenteile

3.2.3.1 Vorlauftabellen, insbesondere Zeitreihentabellen

Viele wichtige Anwendungsbereiche für ein Quaderverfahren mit Berücksichtigung von Schätzintervallen erschließt die Frage nach den Quellen solcher Vorinformationen. Dieses Vorwissen wird sich in aller Regel auf zuvor veröffentlichtes Datenmaterial, auf Vorlauftabellen, gründen.

Für eine Schätzintervallbestimmung geeignete Vorlauftabellen, zu denen auch Zeitreihentabellen gehören (siehe unten), haben oft die gleiche oder eine ähnliche Struktur wie die zu bearbeitende Tabelle. Außerdem sind sie häufig noch nach weiteren Merkmalen ohne Summenbildungen, d.h. nach Parametern von der Art der Zeit angeordnet. Darin gliedert sich dann die aktuelle Tabelle als ein noch nicht veröffentlichter Daten-Teil ein. Obwohl die zu bearbeitende Tabelle mit den anderen Tabellen dieser „Gliederung“ keine Aggregate gemeinsam zu haben braucht und auch keine Summenbeziehungen zur Rückrechnung verfügbar sein müssen, besteht u.U. hier schon aufgrund der Ähnlichkeit hinsichtlich der Ordnungsparameter benachbarter Werte die Gefahr einer Offenlegung geheimer Werte.

Beispielsweise wird der professionelle Nutzer von Zeitreihentabellen, d.h. von Tabellen mit einheitlicher Gliederungsstruktur, die nach festen Zeitabschnitten fortlaufend veröffentlicht werden (Monats-, Vierteljahres-, Jahrestabellen), bereits vor der Veröffentlichung der aktuellen Tabelle über recht genaue Schätzungen der Tabellenwerte verfügen. Dazu gibt es einschlägige Prognoseverfahren (vgl. auch 5.3.2). Er kann damit für jeden Tabellenwert ein Schätzintervall angeben, das bei der Sicherung der aktuellen Zeitreihentabelle gegen zu genaue Rückrechnung geheimer Werte zu berücksichtigen ist. Die notwendige Sicherung geheimer Werte in Zeitreihen betrifft nicht nur die aktuelle Tabelle; sie ist ebenso auch für die zeitlich vorangegangenen Tabellen von Bedeutung, deren Werte aus denen der aktuellen Tabelle ebenfalls geschätzt werden könnten.

In Zeitreihen werden Vorlauftabellen durch Berücksichtigung von Schätzintervallen bei der Sicherung der aktuellen Tabelle weitgehend mitgesichert, weil in älteren Tabellen gesperrte Werte ungenauere aktuelle Schätzwerte

---

8 Bei Vorliegen von Nullwerten gilt dann abweichend von (9) $\varepsilon_Q = \min_{X \in Q \setminus 0} [\min \varepsilon_{min}(X), X_0(0)]$ (9′).

Sei nun $X' = 0 \in Q_u$, $0 \notin Q_g$ in positiver Tabelle.

$\Rightarrow \min_{X' \in Q_u} X' = 0$ und mit (10) $\Rightarrow \varepsilon_{Q+} = 0$ in Übereinstimmung mit ε_+ nach (7a).

Gemäß (10) gilt $\varepsilon_{Q-} = \min [\varepsilon_Q, \min_{X \in Q_g} X] \leq \varepsilon_Q$

mit (9′) und (7b) folgt

$\varepsilon_Q = \min_{X \in Q_g} [\min \varepsilon_{min}(X), \min_{X' \in Q_{u \setminus O}} \varepsilon_{min}(X'), X_0(0)] \leq \min_{X \in Q_g} [\min(X - X_u), \min_{X' \in Q_u \setminus 0}(X_0' - X'), X_0(0) - 0] = \varepsilon_-$, so dass $\varepsilon_{Q-} \leq \varepsilon_-$ ist.

hervorbringen, die auf Grund ihres größeren Schätzintervalls in der laufenden Sicherung als bevorzugte Sperrpositionen gesehen werden. Dadurch wird das Sperrmuster größtenteils über die Tabellen der Zeitreihe durchgereicht und damit gleichzeitig die Rückrechenbarkeit älterer geheimer Werte aus aktuellen Werten weitgehend verhindert. – Ein anderes Sekundärsperrverfahren zum Schutz von Zeitreihentabellen, das auch mit Quaderverfahren ohne jeden Intervallschutz arbeitet, wird im Abschnitt 5.3.2 ausführlich behandelt; es basiert auf der externen Gewichtung von Tabellenwerten. –

Die Anwendung des Quaderverfahrens mit Berücksichtigung extern vorgegebener Schätzintervalle beschränkt sich aber nicht auf zeitlich aufeinanderfolgende gleichartig strukturierte Tabellen, sondern ist mindestens immer dann anzuwenden, wenn der die Daten Veröffentlichende mit anderen bereits veröffentlichten Tabellen für die aktuellen Werte Schätzintervalle berechnen kann!

Darüber hinaus wird man im Allgemeinen aber auch bei Unkenntnis von bereits veröffentlichten, Vorwissen enthaltenden Tabellen pauschalierte Schätzintervalle unterstellen. Erfahrungsgemäß kann man für nicht zu kleine Tabellenwerte ± 50% als Schätzfehler ansetzen. Für die kleinen Werte und Null-Werte empfiehlt sich der Ansatz von Absolutabweichungen, die in ihrer Größe mit denen der kleinsten aus dem relativen Fehleransatz zu berechnenden vergleichbar sein sollten. Damit wird erreicht, dass auch sehr kleine Werte und Nullwerte als Sicherungspartner geheimer Werte in Frage kommen (vgl. 3.2.2.2). Die dazu erforderliche Abgrenzung der als klein zu betrachtenden Tabellenwerte wird man i. Allg. anhand von Testauswertungen vornehmen.

3.2.3.2 Überlappende Tabellen, insbesondere Untertabellen

Eine etwas andere Qualität der Eintragung von Vorwissen in Gestalt von Schätzintervallen liegt bei sog. überlappenden Tabellen vor, d.h. bei Tabellen, die gewisse Aggregate gemeinsam haben. Für die Sicherung geheimer Werte ist dabei notwendig, dass die in mehreren Tabellen gemeinsam vorkommenden Werte den selben Geheimhaltungsstatus besitzen (zur Behandlung überlappender Tabellen siehe auch Punkt 6). Das bedeutet u.a., dass die vorgegebenen Schutzintervalle der in mehreren Tabellen gemeinsam auftretenden Werte in keiner der Tabellen unterschritten werden dürfen. Dabei ist davon auszugehen, dass das tatsächliche Schutzintervall eines geheimen Wertes in einer Tabelle als Schätzintervall in jeder anderen Tabelle, in der er vorkommt, zu berücksichtigen ist. Bei überlappenden Tabellen muss man also die Schutzintervalle der Überlappungswerte in anderen Tabellen als Vorinformation bei der laufenden Bearbeitung in Betracht ziehen.

Das gilt insbesondere auch für die „Übertragung“ von Schutzintervallen beim Untertabellenabgleich mit Intervallschutz. Dazu betrachte man nochmals die Beispieltabelle der Abbildung 3.1 und denke sich diese bezüglich der Spalten weiter untergliedert, wobei jede Spalte (mit Ausnahme der Summenspalte) in zwei weitere Kategorien aufgeteilt wird mit Werten, wie in der Abbildung 3.5 gezeigt. Die Gesamttabelle soll nur als positive Tabelle gesichert werden. Die Eintragungen „p“ bzw. „s“ kennzeichnen die primär- bzw. sekundär geheimen Werte:

Abb. 3.5

100 p	0	100 p	39 s	41	80 s	180
1 s	89	90 s	1 s	0	1 s	91
101	89	190	40	41	81	271

Wenn man in dieser Tabelle zunächst nur die Untertabelle aus den dunkler markierten Spalten betrachtet, so erhält man durch die Sicherung des geheimen Wertes 100 für jeden der Quaderwerte in der Untertabelle höchster Aggregationsstufen die Schutzintervalltabelle gemäß (4).

Abb. 3.6

[99; 180]	[0; 81]	180
[10; 91]	[0; 81]	91
190	81	271

Für die aus den drei ersten Spalten bestehende Untertabelle ergibt sich unter den gleichen Voraussetzungen nicht negativer Tabellenwerte mit (4) bei Berücksichtigung des Aggregationsstufenwechsels innerhalb des Quaders.

Abb. 3.7

[0; 101]	0	[0; 101]
[0; 101]	89	[89; 190]
101	89	190

Der Tabellenwert 100 scheint also in allen Hierarchiestufen der durch Zwischensummen unterteilten Gesamttabelle, Abb. 3.5, hinreichend gesichert, wenn die Untertabellen hinsichtlich der Schutzintervalle als unabhängig voneinander betrachtet werden.

Werden aber die in der Untertabelle höchster Aggregation berechneten Schutzintervalle als externe Schätzintervalle auf die linke Untertabelle unterster Aggregation (die drei ersten Spalten von Abb. 3.5) übertragen, so ergibt sich

Abb. 3.8

X_1	0	$99 \leq X_1 \leq 180$
X_2	89	$10 \leq X_3 \leq 91$
101	89	190

Nach Elimination der Unbekannten X_2 und X_3 erhält man für die Intervallschranken des zu sichernden Wertes X_1 die Abschätzungen

Erste Zeile: $0 \leq X_1$ ^ $99 \leq X_1 \leq 180$

Zweite Zeile: $0 \leq 101 - X_1$ ^ $10 \leq 190 - X_1 \leq 91$

$\Rightarrow$ $99 \leq X_1 \leq 101$

Bei Vernachlässigung der für die ganz linke Untertabelle wirksamen „externen“ Schätzintervalle (in ihrer Summenspalte) hat man mit dem Schutzintervall (0; 101) mit der relativen Spannweite von 101 % einen ausreichenden Intervallschutz, wo hingegen bei Eintragung der Schätzintervalle gemäß Abb. 3.8 ein Schutzintervall (99; 101) mit einer relativen Intervalllänge von 2 % dieser Schutz nur noch dürftig ausfällt!

Diese Betrachtung macht deutlich, dass bei Quadern, deren Werte z.T. auch in anderen Untertabellen vorkommen, die mit (5) gemäß minX‘+ minX berechnete Quaderspannweite die tatsächliche, d.h. aus allen offenen Tabellenwerten zu berechnende Spannweite des zu schützenden Pivots u.U. erheblich überschätzt. Da die Auswahl von Sicherungsquadern aber nach deren Spannweiten erfolgt, haben zu sichernde Pivotelemente mit Schutzsperrungen im Überlappungsbereich von Untertabellen mitunter keinen hinreichenden Intervallschutz!

An dieser Stelle ist allerdings darauf hinzuweisen, dass das Verfahren des (Unter-)Tabellenabgleichs keinen hinreichenden Schutz bieten kann, selbst dann nicht, wenn nur eine genaue Rückrechnung geheimer Werte vermieden werden soll, wenn also auf Intervallschutz ganz verzichtet wird. Erst nach Überführung einer durch Zwischensummen unterteilten Tabelle in eine zwischensummenfreie Tabelle durch Aufstockung der Tabellendimension bietet die Anwendung des Quaderverfahrens einen hinreichenden Intervallschutz – auch bei vorhandener Vorinformation in Gestalt von Schätzintervallen (vergleiche dazu Abschnitt 6.2).

Um die Sicherheit von geheim zu haltenden Werten, die nur mit Quadern im Überlappungsbereich zu schützen sind, wesentlich zu verbessern, sollte beim Abgleich überlappender Tabellen eine Übertragung von Schutzintervallen erfolgen. Das kann wie bei der Behandlung von Schätzintervallen als Vorinformation geschehen, indem diese Schutzintervalle als Schätzintervalle abgespeichert werden. Ist für einen Wert bereits ein Schätzintervall eingetragen, so ist das neue Schätzintervall durch die Schnittmenge aus dem alten Schätzintervall und dem außerdem zu berücksichtigenden Schutzintervall gegeben. Dass hier die Schnittmenge als neues Schätzintervall zu wählen ist, liegt in der Einengung des Wertebereichs, die mit jeder zusätzlichen Information einhergeht, begründet. – Auf diese Weise wird zugleich auch die im Datenbestand noch vor der Auswertung mit dem Quaderverfahren eingebrachte Vorinformation mitberücksichtigt!

Beim fortlaufenden Überschreiben von zuvor bereits eingetragenen Schätzintervallen durch im jeweiligen Bearbeitungsschritt ermittelte noch kleinere Schätzintervallgrenzen besteht die Gefahr des Kollabierens der Schätzintervalle, sodass schon nach wenigen Iterationsschritten des Abgleichsverfahrens wegen zu kleiner Schätzfehler keine Sicherungsquader mehr gefunden werden können.

Das lässt sich bei Tabellen ohne Überlappungsbereiche vermeiden, weil bereits die einmalige Eintragung eines Quaderfehlerwertes ε in das betreffende Feld genügt. Sollte außerdem noch ein anderer Quader den betrachteten Wert mit einem noch kleineren ε-Wert belegen, so ist das für den Intervallschutz irrelevant, weil der externe Tabellennutzer den ε-Wert dann höchstens bis auf den größeren der beiden ε-Werte eingrenzen könnte (vergleiche die Rechtfertigung des Quaderfehlers ε für den Intervallschutz am Ende von 3.1.2 und siehe auch 3.2.3.3, insbesondere die Intervall-Ausgabe-Regel).

Der Tabellenabgleich bietet hingegen keinen hinreichenden (Intervall-)Schutz, so dass dieses Argument nur ein Hinweis auf eine praktikable Verbesserung des Intervallschutzes im Überlappungsfall sein kann: Beim iterativen (Unter-) Tabellenabgleich mit Übertragung von Schutzintervallen als Schätzintervalle wird das bereits im Eingabedatenbestand vorhandene Schätzintervall von GHMITER und bei überlappenden Einzeltabellen auch von QUIT nur einmal durch die Schnittmenge mit einem noch stärker eingrenzenden Intervall überschrieben.

Auf jeden Fall wird man bei der Übertragung von Schätzintervallen beim Tabellenabgleich sowohl den unteren als auch den oberen Abstand des betreffenden Wertes von seinen Schätzintervallgrenzen berücksichtigen müssen: Die Berechnung von Schutzintervallen mit Hilfe der Näherungen (9) bzw. (10) an Stelle der Beziehungen (7a) und (7b) führt insbesondere bei entarteten Schätzintervallen, bei denen eine Intervallgrenze mit dem Wert selbst zusammenfällt oder beinahe zusammenfällt, zu einer zu starken Eingrenzung der Schutzintervalle und damit auch zu vermeidbaren zusätzlichen Sperrungen (Übersperrungen).

Wendet man das Verfahren des iterativen Abgleichs überlappender Tabellen mit Berücksichtigung von Schätzintervallen auf die Beispieltabelle Abb. 3.5 an, so erhält man in der linken Untertabelle niedrigster Aggregation die Summensperrungen in der dritten Zeile, den sekundär geheimen Wert 101, und die Eckfeldsperrung 190. Der zweite sekundär geheime Wert in der Summenspalte bleibt bei der Quaderauswahl zunächst unberücksichtigt, weil der damit gewährleistete Intervallschutz mit einer relativen Spannweite von 2 % zu gering erscheint. – Es sei hier 20 % als nicht zu unterschreitende relative Mindestspannweite angenommen. – Nach diesem Arbeitsschritt hat die

aus den drei linken Spalten der Tabelle Abb. 3.5 bestehende Untertabelle die Gestalt (die neu hinzutretenden Schutzintervalle sind mit * markiert, die „Gegensperrung“ 89 zum Wert 90 ist durch s gekennzeichnet):

Abb. 3.9

100 p [99; 180]*	0	100 p [99; 180]
1	89 s	90 s [10; 91]
101 s [100; 181]*	89 s	190 s [189; 270]*

Die Berechnung der Schutzintervalle für diese Quaderwerte erfolgt zunächst vereinfachend nach (9) und (10) in Verbindung mit den Quaderwerteschätzern (3a) und (3b), wo ε im Intervall $[-\varepsilon_- ; +\varepsilon_+]$ liegt (die Intervallrandwerte miteinbezogen). Dabei ist ferner zu berücksichtigen, dass allen Quaderwerten, die nur die Vorinformation beinhalten, nicht negativ zu sein, das Schätzintervall $[0; \infty)$ zuzuordnen ist; der Wert ε_{min} beträgt für diese Werte also Wert – 0 = Wert. Für ε_{min} ergibt sich im Einzelnen 100 für den Wert 100 des linken oberen Eckfeldes, 1 für den Wert 100 im rechten oberen Eckfeld sowie 101 bzw. 190 für die beiden Summen-Sekundärsperrungen, sodass $\varepsilon_Q = 1$ herauskommt. Da keiner der Quaderwerte kleiner als 1 ist, erhält man mit (10) $\varepsilon_{Q+} = \varepsilon_{Q-} = \varepsilon_Q = 1$ bzw. range = 2; 100 p ist demnach immer noch unzureichend gesichert: Die Abschätzung für ε_{min} ist hier ersichtlich zu stringent, denn bei exakter Abschätzung ergäben sich die Ungleichungen nach Elimination von X_2 und X_3:

Erste Zeile: $0 \le X_1 \quad \wedge \quad 99 \le X_1 \le 180$

Zweite Zeile: $0 \le X_1 + 1 \quad \wedge \quad 0 \le X_1 + 90$

$\Rightarrow \quad 99 \le X_1 \le 180$

$I_1 = [99; 180]$

Durch Anwendung von (7a) und (7b) kommt man zum selben Ergebnis, nämlich keine weiteren Einengungen des Schutzintervalls des primär geheimen Wertes 100: Für das Pivot in der linken oberen Quaderecke existiert nur eine gerade Quaderteilgesamtheit, sodass sich aus (7a) und (7b) Folgendes ergibt:

$\min (X_0 - X) = \min (\infty; 80; \infty; \infty) = 80 \quad \Rightarrow \varepsilon_+ = 80$

$\min (X - X_u) = \min (100; 1; 101; 190) = 1 \quad \Rightarrow \varepsilon_- = 1$

$\varepsilon \in [-1; 80]$

Mit der Quaderschätzwertformel für die gerade indizierten Quaderwerte erhält man mit (3) $\hat{\mathbf{X}} = X + \varepsilon$ die Intervalle [99; 180] für X = 100 in beiden Feldern, [100; 181] für X = 101 und [189; 270] für das Ecksummenfeld.

Um die Sicherung der Gesamttabelle durch iterativen Untertabellenabgleich konsequent weiterzuführen, muss auch die aus der vierten bis sechsten Spalte der Tabelle, Abb. 3.5, bestehende Untertabelle unterster Aggregation mit Übertragung der Schutzintervalle aus höherer Hierarchie bearbeitet werden ([-1; 80] für die ungerade indizierten und [-80; 1] für die gerade indizierten Quaderwerte des Quaders in der Tabelle höchster Aggregation mit Pivot = 100; die Aggregationsstufensumme ist 3 für alle vier Quaderwerte im Inneren der Untertabelle höchster Aggregation!).

Abb. 3.10

39 s [0; 40]*	41	80 s [0; 81] [41; 81]*
1 s	0	1 s [0; 81]
40 s [1; 41]*	41	81 s [42; 82]*

Die Sekundärsperrungen in der mittleren Zeile sind zwar beim ersten Durchlauf gesetzt worden, sie haben für das Weitere aber keine Bedeutung, weil wegen der Sperrung des Ecksummenfeldes in der in Abb. 3.5 linken Untertabelle niedrigster Aggregation auch das Ecksummenfeld in der rechten Untertabelle, der Wert 81, gesperrt werden muss. In der Tabelle 3.10 werden daher die Summenwerte 40 und 81 sekundär gesperrt. Analoge Rechnungen wie bei der Untertabelle Abb. 3.9 liefern (Pivot = 39) $\varepsilon \in$ [-39; 1] und die Schutzintervalle [41; 81] für 80, [0; 40] für 39, [1; 41] für 40 und [42; 82] für das Ecksummenfeld 81.

Der Summenwert 80 in Abb. 3.10 wird demnach durch eine höhere untere Intervallgrenze stärker eingeengt, als bei der vorangegangenen Sicherung in der höheren Hierarchiestufe. In der Untertabelle höchster Aggregation ist nun zu prüfen, ob die Intervalle des neu eingetragenen Sicherungsquaders mit den Schätzintervallen [99; 180] für den primär geheimen Wert 100, [41; 81] für den Wert 80, [42; 82] für 81 und [189; 270] für 190 durch die gemäß (7a), (7b) zu berechnenden Schutzintervalle weiter eingegrenzt werden. Bei der Berechnung der Intervalle ist zu berücksichtigen, dass die das Pivot-Element 100 enthaltenen Spaltenwerte des Quaders wegen der zu addierenden Aggregationsstufen ungerade indiziert sind; das betrifft also die Werte 100 und 190. Die Quaderwerte 80 und 81 sind gerade indiziert.

$\varepsilon_+ = \min [\min (1;1) ; \min (1;1)] = 1; \quad \varepsilon_- = \min [\min (80; 80) ; \min (39; 39)] = 39 \quad \Rightarrow \quad \varepsilon \in [-39; 1]$

Daraus ergeben sich die Schutzintervalle [99; 139] für 100, [189; 229] für 190, [41; 81] für 80 sowie [42; 82] für 81. D.h. der zweite Durchlauf bringt für das primär geheime Tabellenfeld in oberster Aggregation nochmals eine Eingrenzung des Schutzintervalls; die relative Spannweite beträgt jetzt 40 %, beim ersten Durchlauf waren es noch (180-99)/100 = 81 %! Bei Fortführung des Untertabellenabgleichs wären nun die neuen Intervalle zu berücksichtigen, wodurch wieder weitere Eingrenzungen entstehen könnten, die u.U. weitere Sekundärsperrungen erforderten.

Betrachtet man aber die acht gesperrten Quaderwerte, das Karree aus {100; 100; 101; 190} in der linken Untertabelle von Abb. 3.5 und (als darunter liegend) das Karree aus {39; 80; 40; 81} in der rechten Untertabelle unterster Aggregation, als Elemente eines dreidimensionalen Quaders in einer zur vollständigen Tabelle aufgestockten dreidimensionalen Tabelle (zur eingehenderen Darstellung der Aufstockung der Tabellendimension siehe Abschnitt 6.2.2), so sind die diesem Quader zuzuordnenden Schutzintervalle gemäß (4) zu berechnen. Wenn man das oberste linke Tabellenfeld mit dem Wert 100 unterster Aggregationsstufen (1; 1; 1) als Pivotelement wählt, umfasst das zuerst genannte Karree die ungerade indizierten, das zweite die gerade indizierten Quaderwerte (Aggregationsstufen sind zu berücksichtigen). Der kleinste gerade indizierte Quaderwert beträgt demnach $\min X = 39$, der kleinste ungerade indizierte $\min X' = 100$. Die beiden primär geheimen Werte 100 sind daher durch das Intervall [0; 139] mit der relativen Spannweite von 139 % hinreichend geschützt.

Hätte man statt dessen die bereits beim ersten Iterationsschritt erhaltenen Sperrungen in beiden Untertabellen unterster Aggregation, die Werte {100; 100; 1; 90} und {39; 80; 1; 1}, als Elemente eines dreidimensionalen Sicherungsquaders betrachtet, so wäre die ungerade indizierte Quaderteilgesamtheit durch {100; 100; 1; 1} und die gerade indizierte durch {1; 90; 39; 80} gegeben. Die Spannweite dieses Quaders beträgt demnach range = 2; dieser Quader bietet bei der zuvor vorausgesetzten Mindestspannweite von 20 % keinen ausreichenden Intervallschutz für die primär geheimen Werte 100; 100. Er ist also auch im Fall der Geheimhaltung nach dem Dimensionsaufstockungsverfahren zu verwerfen.

Die Betrachtung des Tabellenabgleichs am Beispiel der Untertabellenhierarchie einer kleinen durch Zwischensummen unterteilten Tabelle macht deutlich, dass wiederholtes Abgleichen zu u.U. weit unterschätzten Schutzintervallen führt, womit in der Regel eine deutliche Übersperrung einhergeht. Dies gilt auch dann noch, wenn für die Berechnung der Schutzintervallgrenzen die Beziehungen (7) und nicht deren Näherungen (9) bzw. (10) eingesetzt werden. Dennoch sollte ein Abgleich der Vorinformation in Gestalt von Schätzintervallen berücksichtigt werden, weil die einfachere Betrachtung als positive Tabelle beim Tabellenabgleich ohne Übertragung von Schätzintervallen zu einer deutlichen Überschätzung des damit erzielten Intervallschutzes führen kann – wie es die Behandlung der kleinen Beispieltabelle als vollständige Tabelle zeigt.

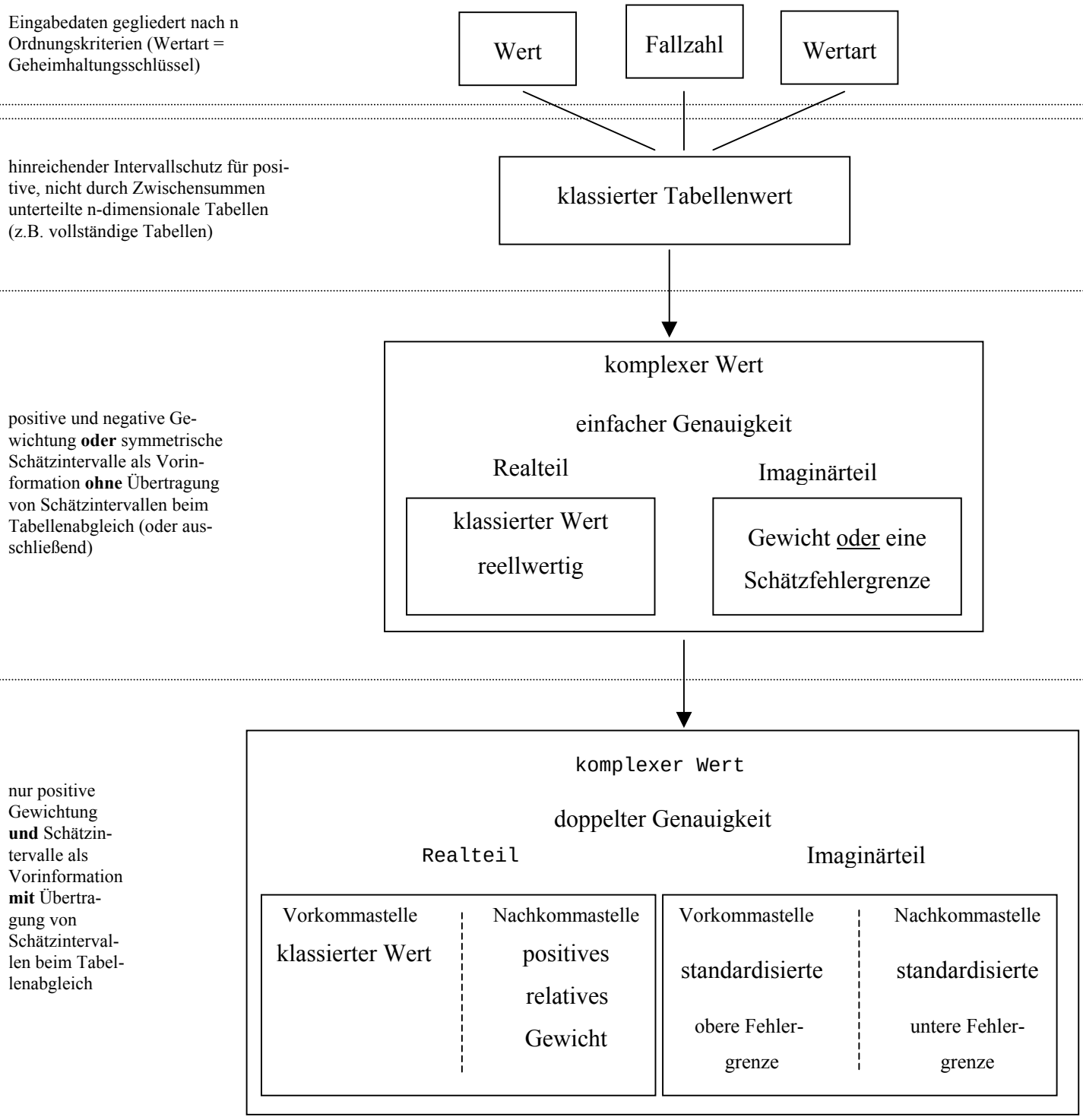
Die Einführung zweier Schätzfehler-Eingabewerte sowie die aus der „statistischen Praxis“ erhobene Forderung, die Schätzintervalle auch im Fall der externen Gewichtung berücksichtigen zu können, erzwingt eine Erweiterung der im Hauptspeicher des Rechners zu führenden Gesamttabelle. Für die Handhabung des nunmehr vier Werte umfassenden ursprünglichen Tabellenwertfeldes der Rechner-Hauptspeichertabelle, die Werteklasse, das Gewicht, der obere und der untere Schätzfehler, ist eine Verschlüsselung in einem „doppelt genauen“ komplexen Wertefeld von Vorteil, weil damit die Gesamtstruktur der Hauptspeichertabelle erhalten bleibt und mit nur einem Zugriff alle vier Werte auf ein Mal erfasst werden können (siehe dazu die nachfolgende „Übersicht zur Struktur des Wertefeldes der Gesamttabelle im Hauptspeicher“)!

Diese Verschlüsselung lässt sich dadurch bewerkstelligen, dass die geeignet standardisierten Einzelwerte jeweils als Wertepaare in je einer der beiden „doppelt genauen“ Festkommagrößen der Komplexen Zahl so gespeichert werden, dass der eine Einzelwert eines Paares vor dem Komma, der andere hinter dem Komma eingetragen wird. Der klassierte Tabellenwert kann z.B. zusammen mit seinem Gewicht im „doppelt genauen“ Realteil angelegt werden, indem der an sich ganzzahlige Klassenwert vor dem Komma und das auf den größtmöglichen Gewichtswert bezogene relative Gewicht nach dem Komma eingetragen wird; der Imaginärteil nimmt dann die beiden standardisierten Schätzfehler auf.

Bei dieser Art der Abspeicherung der vier Werte ist zu beachten, dass das Gewicht – anders als bei der Komplexen Verschlüsselung „einfacher Genauigkeit“ – nur noch positive Werte enthalten kann und keine negativen Gewichte mehr vorkommen, weil das Vorzeichen der Werteklasse mit dem des Gewichts konkurriert; dafür können bei der Quaderauswahl nun Gewichte und Schätzfehler gleichzeitig berücksichtigt und alle Tabellenabgleiche mit Übertragung von Schätzintervallen vorgenommen werden. Da aber der Tabellenabgleich immer nur eine notwendige, keine hinreichende Sicherungsmaßnahme ist, muss bei sehr sensiblen Daten immer die Bearbeitung der zu sichernden Tabelle als vollständige Tabelle angestrebt werden (vergleiche Abschnitt 6.2).

In den beiden zuletzt realisierten EDV-Programm-Versionen zum Quaderverfahren, GHMITER.22 und QUIT, werden Schutzintervalle und Spannweiten im Überlappungsfall bereits mit Übertragung von Schätzintervallen berechnet. Von einer immer noch möglichen Überschätzung des Intervallschutzes sind dabei ausschließlich nur diejenigen Sicherungsfälle betroffen, deren Quader teilweise oder ganz in die Überlappungsbereiche fallen, die anderen sind auch in Bezug auf den Intervallschutz hinreichend gesichert.

Übersicht zur Struktur des Wertfeldes der Gesamttabelle im Hauptspeicher



3.2.3.3 Überlappende Quader, Schutzintervall-Ausgabe anstelle von Schutzsternchen

Die abgespeicherten Schutzintervallgrenzen bieten dem Statistiker eine besonders nutzerfreundliche Behandlung von geheimen Werten in der Veröffentlichungstabelle: Statt der bisher üblichen Unterdrückung geheimer Werte - beispielsweise durch Schutzsternchen - kann er nun die untere und die obere Schutzintervallgrenze des betreffenden Wertes ausgeben. Dies ist allerdings nur dann erlaubt, wenn die zu schützenden geheimen Werte vollständigen Quadern angehören und die nachstehende Intervall-Ausgabe-Regel befolgt wird.

Man braucht dazu zunächst den Begriff des vollständigen Quaders:

Definition:

Ein Quader wird als vollständig bezeichnet, wenn seine Elemente ausschließlich in einer einzigen zwischensummenfreien (Unter-)Tabelle vorkommen (d.h. der Quader ist vollständig, wenn keines seiner Elemente noch in einer anderen Tabelle gesichert wird).

Nur bei der Sicherung geheimer Werte mit vollständigen Quadern kann ihr hinreichender Intervallschutz, mit nicht weiter einzuengenden Schutzintervallgrenzen, garantiert werden. Bei Veröffentlichung von Schutzintervallen hat man dann die folgende Regel zu beherzigen:

Intervall-Ausgabe-Regel:

Bei mehreren vollständigen Quadern gemeinsamen Werten ist für jeden dieser Werte die Vereinigung seiner Schutzintervalle von allen vollständigen Quadern, denen er angehört, als Schutzintervall auszugeben.

Diese Regel ist schon durch Abschnitt 3.1.2 legitimiert. Danach kann der Tabellennutzer kein Schutzintervall innerhalb einer nicht durch Zwischensummen untergliederten (Unter-)Tabelle einengen und dies gilt mit den selben Argumenten auch für den Intervallschutz mit Berücksichtigung von Schätzintervallen: Jeder zu sichernde Tabellenwert wird für sich alleine durch einen vollständigen Quader intervallgeschützt mit festen unverrückbaren Schutzintervallgrenzen. D.h. er hat diesen Schutz unabhängig davon, ob die anderen Tabellenwerte bekannt, näherungsweise bekannt oder selbst schon durch entsprechende Schutzintervalle gesichert sind. Die Ausgabe darf aber keine dieser Schutzintervallgrenzen zu ihren geheimen Werten hin verschieben. Das wird durch die obige Intervall-Ausgabe-Regel sichergestellt.

Die Intervall-Ausgabe-Regel muss bei allen Formen der Schutzintervallausgabe angewendet werden. Dabei ist zu bedenken, dass ein Schutzintervall durch seine Ausgabe zu einem Schätzintervall wird, das beim Tabellennutzer fortan als vorhandene Vorinformation zu unterstellen ist. Das gilt insbesondere auch für die Übertragung von Schutzintervallen auf andere Tabellen, die mit der gerade bearbeiteten gewisse Aggregate gemeinsam haben (vgl. vorhergehenden Abschnitt).

Die Übertragung von Schutzintervallen in den Überlappungsbereich von Tabellen erfolgt durch Schnittmengenbildung der zu Schätzintervallen avancierten Schutzintervalle mit den ursprünglichen Schätzintervallen der Eingabedaten. Die damit verbundene Einengung der ursprünglichen Schätzintervalle steht nicht im Widerspruch zur

Vereinigungsbildung der Schutzintervalle überlappender Quader. Sie entsteht vielmehr durch das Uminterpretieren von den mit dem Quaderansatz berechneten Schutzintervallen in Schätzintervalle, die folglich mit den ursprünglichen, in den Eingabedaten gegebenen Schätzintervallen konkurrieren.

3.2.4 Zusätzliches Wissen: bekannte relative Mindestspannweite

Für den Schutz primär geheimer Werte mittels sekundärer Geheimhaltung ist es äußerst wichtig – und zwar unabhängig vom eingesetzten Sperrverfahren - , dass die für die Auswahl der Sperreintragungen vorgegebene relative Mindestspannweite q dem externen Tabellennutzer keines Falls bekannt gemacht wird! Die Kenntnis der relativen Mindestspannweite ist als eine spezielle Art des Vorwissens zu werten, das, wie jedes Vorwissen über die Tabellenwerte, das Problem der sekundären Geheimhaltung weiter verschärft.

Jedes Sekundärsperrverfahren muss so angelegt sein, dass der externe Tabellennutzer die zu schützenden Tabellenwerte nur bis auf ein Schutzintervall vorgegebener Intervalllänge genau eingrenzen kann. Der für den Schutz geheimer Werte Verantwortliche muss also unterstellen, dass die unteren und oberen Schutzintervallgrenzen X_u , X_o der geheimgehaltenen Tabellenwerte allgemein bekannt sind. (Sie lassen sich jedenfalls mit Hilfe eines Optimierungsansatzes von der Art des in 3.1.1 beschriebenen berechnen; ggf. kann man dabei als Nebenbedingung auch alle Tabellenwerte durch Schätzintervalle eingrenzen, anstatt sie nur als positive Werte zu kennzeichnen.) Kennt der externe Tabellennutzer darüber hinaus noch den Auswahlparameter relative Mindestspannweite q, weiß er also, dass die über die Schutzintervallgrenzen X_u, X_o bekannte Spannweite eines zu schützenden Wertes X größer als das Produkt aus bekannter relativer Mindestspannweite q und unbekanntem geheimen Wert X ist, also $X_o - X_u > q\ X$, so weiß er dadurch, dass der geheime Wert X kleiner als die gegebene Spannweite $X_o - X_u$ dividiert durch die ebenfalls gegebene relative Mindestspannweite sein muss:

$$(X_o - X_u)/q\ >\ X$$

Wenn diese obere Schranke des betreffenden geheimen Wertes X dann auch noch kleiner als seine obere Schutzintervallgrenze X_o ist, hat der externe Tabellennutzer das Schutzintervall $[X_u; X_o]$ von X u. U. stärker eingegrenzt, als es von dem die Tabelle veröffentlichenden Statistiker verantwortet werden kann: Bei bekanntem q ist das Schutzintervall von X gegeben durch

$$X \in [X_u; X_o] \cap [X_u; (X_o - X_u)/q).$$

Eine exakte Offenlegung des zu schützenden Wertes X verbietet die Auswahlregel $(X_o - X_u)/X > q$, denn damit sind ja die Schutzintervallgrenzen X_u, X_o festgelegt worden. Dennoch können sich in für die Geheimhaltung von X besonders ungünstigen Fällen untere und obere Intervallgrenze gemäß $X_u \le X < (X_o - X_u)/q$ beliebig nahe kommen, was einer faktischen Offenlegung von X entspricht.

Beispiel: Seien $X_u = 1$; $X_o = 2$ die Schutzintervallgrenzen eines primär geheimen Wertes und q = 0,999 gegeben, d.h. dem externen Tabellennutzer bekannt, so liegt X im Intervall

$$1\ \le\ X\ <\ (2 - 1)/0{,}999\ \approx\ 1{,}001,$$

ist also bis auf plus minus 0,05 % genau bestimmt. Wäre hier q = 0,5 als Auswahlparameter anzuset-

zen gewesen, hätte der externe Tabellennutzer durch die Kenntnis der relativen Mindestspannweite q praktisch keine Eingrenzung des ursprünglichen Schutzintervalls [1 ;2] von X erreichen können: (2 - 1)/0,5 = 2. Parameterwerte $q \geq 1$ scheiden hier aufgrund obiger Intervallgrenzen $X_u = 1$; $X_o = 2$ wegen der Auswahlregel aus, denn der größte Wert von $(X_o - X_u)/X$ ist 1, dieser Wert soll aber echt größer als q sein.

Anmerkung:

Dass es sich um einen zu schützenden primär geheimen Wert handelt, erkennt der Tabellennutzer oftmals schon an der trotz Wertesperrung veröffentlichten Anzahl der Berichtenden!

Es ist besonders zu betonen, dass diese Betrachtungen an keiner Stelle Bezug auf das Quaderverfahren nehmen; die gemachten Aussagen gelten dem gemäß für alle sekundären Geheimhaltungsverfahren, die ihren Intervallschutz an der Schutzintervalllänge der zu sichernden Werte messen, also auch für das Quaderverfahren für jedwede Art seines Intervallschutzes.

Abschließend sei noch eine kleine Beispieltabelle mit primär und sekundär geheimen Werten angegeben, die mit der Vorinformation, es handelt sich um eine positive Tabelle, bezüglich des primär geheimen Wertes X_1 einen hinreichenden Intervallschutz bietet, wenn der Auswahlparameter relative Mindestspannweite q unbekannt ist. Bei bekanntem q, d.h. bei vorausgesetzter über die Positivität der Tabelle hinausgehender Zusatzinformation "bekannte relative Mindestspannweite" bricht der Intervallschutz vollkommen zusammen:

Abb. 3.11

Kreise \ Wirtschaftsgruppen	A	B	∑
Kreis 1	X_1	X_2	200,0
Kreis 2	X_3	X_4	50,2
Regierungsbezirk	150,1	100,1	250,2

Um dem unbekannten Wert X_1 "näher zu kommen", eliminiert der externe Tabellennutzer wieder die drei restlichen unbekannten X_2, X_3 und X_4 mit Hilfe der Summenbeziehungen der Tabelle und erhält

Abb. 3.12

Kreise \ Wirtschaftsgruppen	A	B	∑
Kreis 1	X_1	200,0 - X_1	200,0
Kreis 2	150,1 – X_1	50,2 – (150,1 – X_1)	50,2
Regierungsbezirk	150,1	100,1	250,2

Wenn dem externen Tabellennutzer keine weiteren Informationen über die Tabellenwerte zugänglich sind, kann er den freien Parameter – bei dieser Auflösung des Gleichungssystems X_1 – nicht weiter eingrenzen. Verfügt der Tabellennutzer aber beispielsweise über das Vorwissen, dass es sich bei dieser Tabelle um eine sogenannte positive Tabelle handelt, so kann er mit diesem Wissen den freien Parameter X_1 weiter eingrenzen:

1. Zeile (Kreis 1): $0 \leq X_1 \leq 200{,}0 \quad \wedge \quad 0 \leq 200{,}0 - X_1 \leq 200{,}0$

d.h. $0 \leq X_1 \leq 200{,}0$

2. Zeile (Kreis 2): $0 \leq 150{,}1 - X_1 \leq 50{,}2 \quad \wedge \quad 0 \leq 50{,}2 - (150{,}1 - X_1) \leq 50{,}2$

d.h. $99{,}9 \leq X_1 \leq 150{,}1$

Die Eingrenzungen sind die Folge der Positivität der Tabelle, wonach keiner der Tabellenwerte die jeweilige Randsumme übersteigt und außerdem jeder Wert größer oder höchstens gleich Null sein kann.

Die Spaltengleichungen liefern keine weitere Eingrenzung von X_1. Wenn dem Tabellennutzer keine weiteren Informationen über die Tabellenwerte vorliegen, ist der Parameter X_1 des o.g. Gleichungssystems zur Berechnung der vier Unbekannten X_1, X_2, X_3, X_4 in obiger Beispieltabelle mit dem Schutzintervall

$$99{,}9 \leq X_1 \leq 150{,}1$$

immer noch hinreichend gesichert.

Wenn dem externen Tabellennutzer aber bekannt ist, dass X_1 primär geheim ist und die relative Mindestspannweite q = 50 % beträgt, so kann er obiges Schutzintervall [99,9; 150,1] von X_1 gemäß $(X_o - X_u)/q = (150{,}1 - 99{,}9)/0{,}5 = 100{,}4 > X_1$ weiter einengen auf das Intervall [99,9; 100,4). X_1 ist dann nicht mehr gesichert, denn es ist wie folgt eingegrenzt

$$99{,}9 \leq X_1 < 100{,}4;$$

seine relative Spannweite ist (100,4 – 99,9)/99,9 = 0,5 %, unterschreitet also die relative Mindestspannweite von q = 50 % drastisch.

Dieses kleine Beispiel macht wiederum deutlich, dass die für die Auswahl von Sperrkandidaten zur Sicherung geheimer Tabellenwerte vorzugebende relative Mindestspannweite selbst vor Offenlegung geschützt werden muss, um die Geheimhaltung von sensiblen Tabellenwerten nicht zu gefährden.

Anmerkungen zu Abschnitt 3.2:

1. Anders als bei der Behandlung von Tabellen mit alleiniger Berücksichtigung der Vorinformation „positive Tabelle“ tritt durch die Eingrenzung der Tabellenwerte durch Schätzintervalle u.U. ein Widerspruch zur geforderten Intervallsicherung auf, nämlich mindestens immer dann, wenn das von außen angegebene Schätzintervall kleiner ist als das zu seinem Schutz vorgegebene. Ein Tabellenwert, der dem Tabellennutzer bereits bis auf wenige Prozent bekannt ist, lässt sich eben mit keinem Sicherungsverfahren der Welt durch Sekundärsperrungen noch offener Tabellenwerte so schützen, dass danach Angaben über diesen Wert nur noch im Bereich von beispielsweise 100 % Abweichung möglich sind.

2. Von praktischer Bedeutung ist auch, dass mit der Berücksichtigung von Schätzintervallen nicht mehr alle offenen Werte als Sperrkandidaten in Betracht kommen, weil ihre Schätzintervalle sie zu genau eingrenzen: Durch Berücksichtigung von Schätzintervallen kann somit verhindert werden, dass der Tabellennutzer auf Grund der u.U. von ihm vorzunehmenden engen Eingrenzung von Tabellenwerten durch sein Vorwissen primär geheime Werte zu genau berechnen kann. Umgekehrt kann die Angabe sehr kleiner Schätzintervalle aber auch dazu genutzt werden, gezielt bestimmte Tabellenwerte von der Sekundärsperrung von vorneherein auszuschließen; man darf von solchen Maßnahmen aber nur sparsam Gebrauch machen, weil sonst wegen fehlender Sicherungspartner die Sicherung der ganzen Tabelle auf dem Spiel steht.

3. Anders als bei der Behandlung von Tabellen mit sowohl positiven als auch negativen Werten ohne Berücksichtigung von Schätzintervallen tritt durch die externe Eingrenzung der Tabellenwerte auch bei „nicht positiven“ Tabellen das Problem des Intervallschutzes auf. Während bei Tabellen ganz ohne Vorinformation, d.h. auch ohne die Vorinformation, dass es sich um eine positive Tabelle handelt, keinerlei Beschränkung des Parameters ε der Quadergleichungslösung zu berücksichtigen ist, wird bei vorhandener Vorinformation, „Schätzintervalle“, die Auswahl von Sicherungsquadern auch bei Tabellen mit positiven und negativen Werten eingegrenzt, ganz ähnlich, wie bei positiven Tabellen.

Teil II: Erweiterungen und Anwendungen

4. Verallgemeinerung des Quadermodells

4.1 Quaderverfahren zur Werteverfälschung

Das Quadermodell zur Sicherung geheimer Werte ist nicht prinzipiell nur auf das Sperren von Tabellenwerten zugeschnitten. Es kann auch für die unterschiedlichen Formen der Geheimhaltung durch Werteverfälschen eingesetzt werden, sei es, dass die Werte jedes Quaders innerhalb der ranges durch Zufallszahlen modifiziert werden oder, dass die Verfälschung durch Umbuchungen erfolgt (vergleiche G. Appel, Dublin 1992), wenn jedenfalls an der Forderung, höhere Aggregate weitgehend zu verschonen, festgehalten wird.

Bei Einsatz des Quaderverfahrens zur Werteverfälschung durch Überlagerung von Zufallsfehlern muss man berücksichtigen, dass der zufällig ausgewählte Fehler ε für alle Werte eines betrachteten Sicherungsquaders dem Betrage nach gleich, aber mit unterschiedlichem Vorzeichen für die beiden Quaderteilgesamtheiten gewählt werden muss (vergleiche Abschnitt 3, Gleichung 3), damit die Randsummen der entsprechenden Untertabelle unverändert bleiben. Damit diese Überlagerung darüber hinaus nicht zu Unverträglichkeiten mit der für die beabsichtigte Quadersicherung zu Grunde gelegten Vorinformation führt, muss ε in einem Fehlerintervall

$$\varepsilon \in [-\varepsilon_{-} , +\varepsilon_{+}]$$

liegen. Dabei bestimmen sich die Intervallgrenzen $-\varepsilon_{-}$, ε_{+} aus der Vorinformation: Bei Vorgabe von Schätzintervallen gelten die Beziehungen (7a, b) bzw. deren Näherungen (10). Wird nur die Positivität der Tabelle vorausgesetzt, ergibt sich die untere Fehlergrenze als negativer kleinster gerade indizierter Wert und die obere Grenze als positiver kleinster ungerade indizierter Wert des betreffenden Sicherungsquaders (vergleiche (4), Abschnitt 3.1.2). In beiden Fällen ist ε mit einem positiven Vorzeichen zu versehen, wenn der zu verfälschende Quaderwert gerade indiziert ist und mit einem Minuszeichen, wenn der Quaderwert der ungerade indizierten Teilgesamtheit angehört (Aggregationsstufenwechsel werden dabei mitberücksichtigt).

Zur Verfälschung der Werte eines Quaders wird also ein beliebiger Fehler ε aus dem oben angegebenen Fehlerintervall zufällig ausgewählt und dann zu allen gerade indizierten Quaderwerten addiert und von allen ungerade indizierten subtrahiert. Die Zufallsauswahl der ε kann mit Hilfe eines Zufallszahlengenerators geschehen, der gleichverteilte Zufallszahlen liefert. Ggf. lassen sich auch Zufallszahlengeneratoren zur Erzeugung normalverteilter oder nach anderen Verteilungsfunktionen verteilter Zufallszahlen verwenden.

Bei der solchermaßen modifizierten Tabelle bleiben z.B. die vormals einzelnen Berichtenden zugeordneten Werte zwar verfälschte, aber weiterhin doch Einzelangaben. Das kann hier ein einfaches Umbuchungsverfahren ändern,

bei dem die Fallzahlen zwischen jeweils benachbarten Quaderwerten (also zwischen Quaderwerten, die zur selben Randsumme beitragen) so ausgetauscht werden, dass alle Quaderwerte mit (beispielsweise) mindestens drei Berichtenden belegt sind. Da dieser Austausch innerhalb eines Quaders erfolgt, kann man damit erreichen, dass die Randsummenwerte der Fallzahlen unverändert bleiben.

Dazu genügt es, die auszutauschende Fallzahl m nach genau demselben Muster wie die zu überlagernde Fehlergröße ε auf die einzelnen Quaderwerte zu verteilen, wobei - zunächst noch abweichend von der Fehlergrenzenbestimmung - zu berücksichtigen ist, dass durch den Umbuchungsprozess keine der Fallzahlen des Quaders kleiner als (beispielsweise) drei werden darf. Zwischen benachbarten Quaderwerten unterschiedlicher Aggregation darf keine Umbuchung erfolgen, beide Fallzahlen erhalten den gleichen Zu- oder Abschlag. Dem gemäß beziehen sich Umbuchungsvorgänge immer nur auf im Quader benachbarte Werte gleicher Aggregation; das Vorzeichen der Fallzahländerungen richtet sich – wie das der Quaderwertverfälschungen – genau nach der Art der Quaderwerte-Indizierung bei Berücksichtigung von Aggregationswechseln.

Allgemein formuliert sei M_o die zulässige kleinste Fallzahl eines noch offenen Tabellenwertes und M_g die Fallzahl des zu sichernden Wertes, dann gelten für die Fallzahlen M eines n-dimensionalen Sicherungsquaders die zu den Ungleichungen (3) analogen Beziehungen (der zu sichernde Wert sei gerade indiziert)

$$M + m \geq M_o \quad \text{und} \quad M' - m \geq M_o \ ,$$

wobei wieder die ungestrichenen Größen gerade, die gestrichenen ungerade indiziert sind. Betrachtet man nun anstelle der Fallzahlen selbst deren reduzierte Größen, die sich durch Subtraktion der kleinsten zulässigen Fallzahl M_o von allen Fallzahlen des betreffenden Quaders ergeben, so gelten die Ungleichungen (3) auch für diese reduzierten Quaderfallzahlen. Die auszutauschenden Fallzahlen besitzen daher die gleiche Struktur wie die Quaderfehler ε einer positiven Tabelle; selbstverständlich müssen sie außerdem auch noch ganzzahlig sein. Demnach ist der größte dem gerade indizierten zu sichernden Wert M_g noch zuzuschlagende Fallzahlwert

$$m_{max} = \min (M' - M_o) \qquad (12) \ .$$

Da die auf das zu schützende Tabellenfeld umzubuchende Fallzahl m durch die Bedingungen festgelegt ist, dass die neue Fallzahl des zu schützenden Feldes nicht kleiner als M_o sein darf, ergibt sich aus $M_g + m_{max} \geq M_o$ und obiger Gleichung die Quaderauswahlbedingung

$$\min (M' - M_o) \ \geq \ (M_o - M_g) \qquad (13) \ ,$$

die zusätzlich zu den Auswahlkriterien zu berücksichtigen ist. Das bedeutet, dass ein Sicherungsquader bei Umbuchung so ausgewählt werden muss, dass die kleinste ungerade indizierte Fallzahl des Quaders nach Abzug der kleinsten zulässigen Merkmalsträgerzahl nicht kleiner ist als die auf das geheime Feld zu übertragende (umzubuchende) Fallzahl. Die kleinste ungerade indizierte Fallzahl des auszuwählenden Quaders muss damit so groß sein, dass sie durch die vorzunehmende Umbuchung nicht selber unzulässig klein wird. Dass hier nur der kleinste

ungerade indizierte Fallzahlwert limitierend wirkt und nicht auch der kleinste gerade indizierte, liegt daran, dass die Umbuchungen wegen der vorausgesetzten geraden Indizierung des Pivots immer nur von den ungerade indizierten Fallzahlen zu den gerade indizierten erfolgen und niemals umgekehrt.

Ein anderer, besonders für die Akzeptanz der zu veröffentlichenden Tabelle ganz wesentlicher Aspekt bei der Auswahl von Sicherungsquadern ergibt sich aus der Forderung der Tabellennutzer, die "Sperrkandidaten" nach Möglichkeit benachbarten Gliederungskategorien zu entnehmen: Beim Austausch von Meldeeinheiten werden die Einheiten selbst umbenannt also verfälscht, wobei diese Verfälschung umso krasser ausfällt, je mehr sich die neue Kategorie, in die die Meldeeinheit umgebucht wird, von der alten, aus der sie stammt, unterscheidet. Wenn man davon ausgehen kann, dass einander sehr ähnliche Kategorien auch in der Tabellengliederung entsprechend nahe beieinander liegen, läuft obige Forderung darauf hinaus, zum Schutz eines geheimen Wertes durch Umbuchung solche Quader auszuwählen, deren Quaderwerte besonders kleine Abstände vom zu sichernden Pivot haben. Die Bevorzugung von einem zu sichernden Wert abstandsmäßig nahe benachbarter Werte als "Sperrkandidaten" kann mit Hilfe des Summenkriteriums durch eine instantane, d.h. während des Sperrvorgangs vorgenommene Gewichtung geschehen, wobei sich die Gewichte aus den Abständen der Quaderwerte vom Pivot-Element berechnen lassen (Näheres dazu unter 5.3.3).

Die Kombination von Werteverfälschung durch Fehlerüberlagerung und Umbuchung von Fallzahlen ist eine einfache Erweiterung des bisher praktizierten Quaderverfahrens mit Intervallschutz, die mit wenigen zusätzlichen Programmbefehlen zu realisieren wäre. Eine wesentliche Vergrößerung des Rechenzeitaufwandes ist dabei nicht zu erwarten. Wesentlich aufwendiger gestaltete sich die quaderweise Umbuchung von Fallzahlen, wenn dabei die zugehörigen gemeldeten Einzelangaben mit übertragen werden sollten. Dann wären nämlich anstelle des Tabellenwertes, der die Summe aus den in das betreffende Tabellenfeld eingetragenen Fallzahlen zugeordneten Einzelangaben darstellt, die Einzelangaben selbst einzufügen, damit der gewünschte Teil von ihnen umgebucht werden könnte. Der zu erwartende Umstellungs- und Rechenzeitaufwand wäre in diesem Fall erheblich.

Andererseits weist das zuletzt angedeutete Umbuchungsverfahren der Übertragung von Fallzahlen mit ihren Einzelangaben auf andere Tabellenfelder den erheblichen Mangel auf, dass die Wertesummen in den Tabellenrändern nicht mehr erhalten bleiben. Lediglich die Fallzahlen werden in den Tabellenrändern durch die Quaderumbuchungen unversehrt gelassen. Die Verfälschungs- und Umbuchungsverfahren wurden bisher nicht gefordert und daher auch EDV-mäßig nicht realisiert.

Solche Betrachtungen verdeutlichen auch den (un-)wesentlichen Unterschied zwischen dem hier diskutierten Sperrverfahren und einem Perturbationsverfahren, das Randsummen nach Möglichkeit unverändert lässt: Beim Sperren bleibt es dem Anwender selbst überlassen, die gesperrten Werte mit einem seinen Anforderungen gerecht werdenden Schätzverfahren zu bestimmen, während bei Perturbation diese Werte bereits vorgegeben sind, z.B. durch die tatsächlichen Tabellenwerte mit überlagerten Zufallsfehlern.

4.2 Quaderverfahren zum Intervallschutz auf Mikrodatenebene

Jede Sicherungsmaßnahme zum Schutze von sensiblen Tabellendaten, primäre wie sekundäre, zielt letztendlich auf die Vermeidung einer zu genauen Offenlegung der zu den Tabellenwerten beitragenden Einzelangaben ab. Diese Rückbesinnung auf das eigentliche Ziel der Geheimhaltung in Tabellendaten führt unmittelbar zu der Forderung, in erster Linie Intervallschutz für das Einzeldatenmaterial anzustreben und nicht nur für deren Aggregate. Da die nachfolgenden Ausführungen u.A. auch die Dominanzproblematik betreffen, werden hier ausschließlich positive Tabellen betrachtet.

4.2.1 Veröffentlichungstabelle mit Mikrodatengliederung

Als Lösungsansatz bietet sich die Doppelquadersicherung von Mikrodaten mit Intervallschutz an, die den zu veröffentlichenden aggregierten Daten in Gestalt einer zusätzlichen Gliederung durch Aufteilen der Tabellenwerte in ihre Einzelwerte angefügt werden. Die Summentabelle dieser Gliederung ist dann die n-dimensionale Veröffentlichungstabelle. Man erhält auf diese Weise eine n+1-dimensionale Gesamttabelle bestehend aus $m \geq 3$ Tabellen von „Einzelangaben“, die alle genau so gegliedert sind wie die Veröffentlichungstabelle, und als Summentabelle die Veröffentlichungstabelle selbst.

Die Gesamtheit der Gliederungskategorien zur Aufteilung der Veröffentlichungstabelle in ihre nicht zu veröffentlichenden Einzelangabentabellen (die durchaus auch aggregierte Werte enthalten, weil sie hinsichtlich der n Gliederungen dieselbe Summenstruktur wie die Veröffentlichungstabelle haben) umfasst mindestens drei n-dimensionale Einzelangabentabellen. Bei weniger als drei Einzelangabentabellen verursachte jede auf unterstem Aggregationsniveau durch Einzelangaben erzwungene Doppelquadersicherung Sperrungen in die gemeinsame Summentabelle, die Veröffentlichungstabelle, und das immer auch dann, wenn mehr als zwei Einzelangaben zu einem Aggregat der Veröffentlichungstabelle beitragen. Drastische Übersperrungen wären die Folge.

Die Wahrung der Geheimhaltung erfolgt nun in der n+1-dimensionalen Gesamttabelle und zwar so, dass zunächst alle Werte der – ohnehin nicht zu veröffentlichenden - Einzelangaben-Tabellen als primär geheim ausgewiesen werden. Ansonsten sind keine weiteren Primärsperrungen erforderlich; insbesondere kann die n-dimensionale Veröffentlichungstabelle völlig frei von Primärsperrungen bleiben, die für die Sicherung der Einzelangaben erforderlichen Sperrungen in der Veröffentlichungstabelle werden durch die Doppelquadersicherung der n+1-dimensionalen Gesamttabelle aus den Einzelangabentabellen eingetragen.

Für den Anwender der EDV-Programme GHMITER und QUIT ist hier anzumerken, dass bei ausschließlicher Primärsperrung von Einzelangaben anstelle der Primärsperrung aller Mikrodaten drastische Übersperrungen hinzunehmen wären: Bei der Sicherung von Untertabellen mit nur zwei Einzelangaben bezüglich jeder Gliederung im Tabelleninneren wird aufgrund der negativen Einzelangabengewichtung durch Werteklassierung immer zuerst der – einzige – Quader im Inneren mit lauter Einzelwerten aufgesucht. Der einzig zu akzeptierende zweite Sicherungsquader kann dann nur noch über das Randsummen-Eckfeld gehen, weil jeder andere Quader außer

dem Pivot noch eine weitere Einzelangabe im Tabelleninneren enthielte, die aber bereits dem ersten Quader angehört und darum nicht in Frage kommt. Würden nun allen Tabellenwerte der Mikrodatengliederung primär gesperrt, stünden für die Doppelquaderbildung genügend gesperrte Randsummen zur Auswahl, auf die das Programm gerne ausweichen würde, weil gesperrte Werte mit mehr als einem Berichtenden einer Werteklasse mit noch kleineren Werten angehören als die Einzelangaben (vgl. 5.2.2).

Andererseits dürfen Mikrodaten-Summen – genau wie Einzelangaben im Rand - selbst nicht gesichert werden müssen, weil sonst ebenfalls Übersperrungen entstünden.(vgl. Abb. 4.1)**.** Letzteres lässt sich erreichen, wenn man bereits im Eingabebestand alle Daten der Mikrodatengliederung als Einzelangaben ausweist und dann alle diese „Einzelangaben“ auch als solche primär sperrt. Dann ist nämlich keine besondere Gewichtung der „Einzelangaben“ in den Mikrodaten-Summen erforderlich, damit diese – wie zuvor gefordert - bei der Quaderauswahl bevorzugt werden, weil die Randsummeneinzelangaben gemäß 2.1.1 nach der Einzelangaben-Randsummen-Regel zu bearbeiten sind. Nach dieser Regel ist eine Einzelangabe im Summenrand wie eine primär geheime Angabe mit mehr als einem Merkmalsträger zu behandeln, die selbst nicht mehr gesichert werden muss.

Die Doppelquadersicherung ist erforderlich, weil alle als Sicherungspartner niedrigster Aggregation im Einzeldatenmaterial in Frage kommenden Tabellenwerte Einzelangaben sind (vgl. 2.1.2). Lassen sich mit diesen Einzelwerten und deren Summen und Zwischensummen keine Doppelquader bilden, muss man in die zu veröffentlichende Summentabelle ausweichen. Auf diese Weise entstehen in der Veröffentlichungstabelle Sperreintragungen, die Primärsperrungen der ursprünglich als n-dimensionale Tabelle zu behandelnden Summentabelle und die zugehörigen Sekundärsperrungen.

Bei Bearbeitung der n+1-dimensionalen Gesamttabelle kann man in der Veröffentlichungstabelle nicht mehr zwischen primärer und sekundärer Geheimhaltung unterscheiden: Aus Sicht der n+1-dimensionalen Gesamttabelle gibt es in der zu veröffentlichenden Summentabelle nur noch Sekundärsperreintragungen. Durch dieses Vorgehen wird die primäre Geheimhaltung bis auf das Sperren auf der ohnehin unveröffentlichten Mikrodatenebene direkt in das Quaderverfahren integriert und zwar ohne dieses in irgend einer Weise verändern zu müssen!

Zur Einsparung von Rechenzeit und Hauptspeicherplatz sollten die Einzelangaben innerhalb der Tabellenfelder der Veröffentlichungstabelle der Größe nach sortiert werden. Man erhält so eine Schar von m Einzelmaterialtabellen, von denen die erste den in jedem Tabellenfeld größten Einzelwert aufweist, die zweite den in jedem Tabellenfeld zweitgrößten Wert etc.. Bei dieser Sortierung genügt bereits die Einbeziehung nur der drei Einzelmaterialtabellen mit den größten Einzelangaben, die anderen sind dann für die Auswahl von Doppelquadern entbehrlich, wodurch eine erhebliche Reduktion des Umfanges der n+1-dimensionalen Gesamttabelle erreicht wird .

Im Folgenden wird nun zur Herleitung eines Mikrodatensicherungskonzepts zunächst von dieser reduzierten n+1-dimensionalen Tabelle ausgegangen, die in Bezug auf die neu hinzugenommene Gliederung nach Einzeldaten zusammen mit der Veröffentlichungstabelle nunmehr vier n-dimensionale Tabellen umfasst, also das vierfache Volumen der ursprünglichen Veröffentlichungstabelle aufweist.

Zur Gewährleistung eines ausreichenden Intervallschutzes für alle Einzelangaben der erweiterten n+1-dimensionalen Tabelle mit Hilfe des Quaderverfahrens kann man im einfachsten Fall die Positivität der Tabelle als Vorinformation unterstellen und außerdem eine geeignete relative Mindestspannweite q vorgeben. Weil es dabei primär nicht um den Schutz bezüglich irgendwelcher Dominanzmaße geht, sondern nur darum, ausreichend große Schutzintervalle für alle Einzelangaben zu erreichen, kann man sich mit q-Werten kleiner als Eins begnügen – ein ganz wesentliches Argument für die Einführung der Einzelwertegliederung! –.

Geht man mit diesen Voraussetzungen die Sicherung von Tabellendaten an, so zeigt sich, dass Dominanzfälle, insbesondere in den höheren Hierarchien der Aggregation, weitgehend unberücksichtigt bleiben. Das verwundert aber nicht, denn es wurde ja kein Vorwissen über die Verhältnisse der Einzelangaben zueinander berücksichtigt, sondern einzig und allein das Vorwissen, dass es sich um eine positive Tabelle handelt. Durch das Wissen aber, dass ein Aggregat positiv ist, weiß man eben noch nicht, ob darin z.B. ein dominierender Wert vorkommt.

4.2.2 Dominanzschutz durch Intervallschutz von Mikrodaten

Dass Dominanzschutz ohne Unterstellung der Kenntnis von Schätzintervallen sinnlos ist, zeigt die folgende Beispieltabelle: Zu sichern sei eine eindimensionale Veröffentlichungstabelle, die unterste Zeile der Beispieltabelle. Darüber sind die gefüllten Einzelangaben-Tabellen angeordnet (es gibt nur zwei davon, die dritte ist leer und wurde daher nicht aufgeführt). In der ersten Zeile jedes Tabellenfeldes steht der Wert, in der zweiten die Anzahl der Berichtenden. Alle Tabellenwerte der durch die Mikrodatengliederung neu hinzugenommenen nicht zu veröffentlichenden Tabellen sind primär geheim; sonst wurden keine weiteren Primärsperrungen gesetzt.

Wird als Vorwissen nur die Positivität der Tabelle unterstellt, so ergibt sich nach Anwendung des Quaderverfahrens bei einer relativen Mindestspannweite von q = 50 % folgendes Muster sekundärer Sperrvermerke (p = primär, s = sekundär geheim):

Abb. 4.1

Einzelwerte \ Gliederung	A	B	∑
größte Einzelwerte	1000 p 1	3 p 1	1003 p 2
zweitgrößte Einzelwerte	2 p 1	1 p 1	3 p 2
Veröffentlichungs-Tabelle	1002 **s** 2	4 **s** 2	1006 4

Dass bei Doppelquadersicherungen nach obigem Muster immer sogar mindestens eine 100 %-Spannweite für jeden Quader herauskommt, liegt daran, dass jeder Quader eine Randsumme mit einbezieht und damit der Minimalwert einer Teilgesamtheit immer der zu schützende Pivot-Wert selbst ist.

Man sieht, dass nach der Bearbeitung der als positiv vorausgesetzten Tabelle nur noch das Summenfeld der Veröffentlichungstabelle, das Eckfeld in obiger zweidimensionalen Tabelle, offen bleibt. Zu diesem Summenfeld tragen aber mit Ausnahme des Wertes 1000 nur sehr kleine Werte bei: Der Anteil des größten (dominierenden) Wertes an der Eckfeldsumme beträgt 1000/1006 = 99,4 %! Das auf den größten Einzelwert bezogene Restaggregat, das man nach Subtraktion der beiden größten Einzelwerte von der Eckfeldsumme erhält, ist (1006 – 1000 – 3)/1000 = 0,3 %. D.h. nach allen bisher üblichen Konzentrationsmaß-Parametern der Einfachdominanzregel oder der sogenannten p%-Regel ist die Eckfeldsumme zu sperren (vgl. auch Abschnitt 0.2).

Ob ein Tabellenfeld aus Dominanzgründen nach einer (n, k)-Dominanzregel oder nach der p%-Regel zu sperren ist und wie groß dabei die betreffenden Parameter zu sein haben, entscheidet der Statistiker aus einem Gefühl heraus. Diese Subjektivität manifestiert sich besonders deutlich durch die breite Palette von Parameterwerten für diese Regeln, für die sich die Sicherheit von Einzeldaten gewährleistenden Statistiker bisher entschieden haben! Was aber letztendlich zu der Aussage, „es handelt sich um einen Dominanzfall“, führt, ist doch die Unterstellung von Vorwissen beim Tabellennutzer, wonach dieser die Einzelwerte durch Schätzintervalle eingrenzen und erst dadurch eine die Größenverhältnisse bestimmende Dominanzangabe machen kann.

In obiger Beispieltabelle wird man als allgemein bekannt unterstellen, dass die kleinen Einzelangaben nicht größer als 10 sein können, während für alle großen Werte, beispielsweise den dominierenden Wert 1000 und den Veröffentlichungswert 1002 ein Schätzintervall von 500 bis 2000 angegeben werden kann – noch bevor die Veröffentlichungstabelle, die untere Zeile der Tabelle (Abb. 4.1), publiziert wird -. Der bei Sicherung der nur als positiv vorausgesetzten Gesamttabelle bei q = 50% hinreichende Quader zum Pivot 1000 mit Diametralwert 4 in der Veröffentlichungstabelle reicht dann aber für den Schutz des Pivots 1000 nicht mehr aus, es muss die Eckwertsumme als Diametralwert genommen werden.

Um sich davon zu überzeugen, berechnet man die Spannweite für den zu kritisierenden Quader Q(1000; 4), d.h. für den Quader mit Pivot 1000 und Diametralwert 4 mit Hilfe der Spannweitenformel für vorgegebene Schätzintervalle gemäß 3.2.1, Formel (7a), (7b):

Abweichung/Werte	X = 1000	X' = 3	X = 1002	X' = 4
$X_0 - X$	2000-1000 = 1000		2000-1002 = 998	
$X' - X'_u$		3-0 = 3		4-0 = 4
$X'_0 - X'$		10-3 = 7		10-4 = 6
$X - X_u$	1000-500 = 500		1002-500 = 502	

$$\varepsilon_+ = \min[\min(1000;998);\ \min(3;4)] = 3; \quad \varepsilon_- = \min[\min(7;6);\ \min(500;502)] = 6$$

$$\text{range} / X_{pivot} = (\varepsilon_+ + \varepsilon_-) / X_{pivot} = (3 + 6) / 1000 = 0{,}9\ \% < q = 50\ \%$$

Da die Sicherungsmöglichkeit mit dem Quader Q(1000; 4) bei unterstellter Vorinformation obiger Schätzintervalle wegen $q = 50\ \%$ ausfällt, bleibt für die Sicherung des Wertes 1000 nur die Sperrung über das Eckfeld, also der Quader Q(1000; 1006), der Quader mit Pivot 1000 und Diametralwert 1006. D.h. die Information über die Kleinheit der drei nicht dominierenden Angaben gegenüber dem dominierenden Wert, die letztlich zur Erkennung eines Dominanzfalls im Eckfeld führt, wird über die Schätzintervalle der Einzelangaben eingetragen, die ein Tabellennutzer schon vor der Veröffentlichung der Tabelle haben kann.

(min kann weggelassen werden)

4.2.3 Begründung der p%-Regel mit dem Quaderverfahren

Weiterführende Untersuchungen des LDS NRW haben ergeben, dass für den Dominanzschutz der gegebenen Statistiktabelle eine **Gliederung nach Einzeldaten** anzufügen ist, die in Bezug auf die unterste Aggregation der gegebenen n-dimensionalen Tabelle **den größten und den zweitgrößten Einzelwert sowie das Restaggregat**, den Tabellenwert abzüglich der beiden größten Einzelwerte, enthalten sollte. Das Hinzufügen weiterer Einzelangaben anstelle des Restaggregats würde zu Übersperrungen führen, weil zur Auswahl eines Doppelquaders zum Schutze des größten Einzelwertes der drittgrößte Einzelwert zum selben Aggregat oft schon zu klein ist, während das Restaggregat als Summe vieler kleiner Werte noch einen ausreichenden Schutz bieten kann.

Mit dieser neuen Gliederung erhält man wieder eine n+1-dimensionale Gesamttabelle, die drei neu hinzutretenden Tabellen des „Einzelmaterials“, von denen jede genau so gegliedert ist wie die gegebene n-dimensionale Tabelle und die gegebene Tabelle als Summentabelle in Bezug auf diese neu eingefügte Einzeldatengliederung. Für den Dominanzschutz ist es unerlässlich, außerdem noch die Schätzfehler anzugeben, die der Tabellennutzer aufgrund des zu unterstellenden Vorwissens bestimmen kann. Um diese n+1-dimensionale Gesamttabelle mit dem Quaderverfahren sichern zu können, müssen noch alle Angaben der drei Einzelmaterialtabellen primär gesperrt werden (vgl. Hinweis zu GHMITER und QUIT unter 4.2.1). Die Anwendung des Quaderverfahrens mit einem die Schätzintervalle der Tabellenwerte berücksichtigenden Intervallschutz liefert dann außer dem Schutz gegen zu genaues Rückrechnen der Einzelwerte auch einen hinreichenden Dominanzschutz.

4.2.3.1 Relative Schätzfehler bis einschließlich 100%

Unterstellt man, dass der Tabellennutzer „seine“ Tabellenwerte mit einem für alle Tabellenwerte einheitlichen relativen Schätzfehler von $\pm f$ mit $0 < f \leq 100\%$ eingrenzen kann, so werden alle Werte der (n-dimensionalen) Veröffentlichungstabelle gesperrt, die nach der p%-Regel geheim zu halten gewesen wären. Dabei ergibt sich der Sicherungsparameter p der p%-Regel als Quotient aus der relativen Mindestspannweite q, nach der die Sicherungsquader auszuwählen sind, und der relativen Schätzintervalllänge 2f, dem zweifachen relativen Schätzfehler f.

$$p = q / (2f), \quad 0 < f \leq 1 \qquad (14)$$

Bei vorgegebener relativer Schätzintervalllänge 2f erfordert demnach ein hoher Sicherheitsanspruch, ausgedrückt durch einen hohen Parameterwert p, auch einen hohen Intervallschutz durch die Quadersicherung, ausgedrückt

durch eine große relative Mindestspannweite q. Muss ein sehr genaues Wissen über die Einzelwerte, d.h. eine kleine Schätzintervalllänge 2f beim Tabellennutzer unterstellt werden, so bedeutet dies bei festgehaltener relativer Mindestspannweite einen hohen Schutzanspruch, ausgedrückt durch einen großen Parameterwert p. Schließlich nimmt bei konstantem p die relative Schutzintervalllänge (in Gestalt der relativen Mindestspannweite q) mit zunehmendem Vorwissen, d.h. mit kleiner werdendem f ab. Mit anderen Worten, **der Schutzanspruch, ausgedrückt durch den Parameter der p%-Regel, gibt nur das Verhältnis von Schutz- zu Schätzintervalllänge an und sagt alleine nichts aus über den tatsächlich gewährleisteten Intervallschutz der Einzelwerte!**

Bei Bearbeitung der n+1-dimensionalen Gesamttabelle setzt das Quaderverfahren außer den mit der p%-Regel mit Parameter p gemäß (14) zu begründenden Sperreintragungen noch weitere Sperrungen, die als Sekundärsperrungen zur Sicherung der p%-Regel anzusehen sind. Auf diese Weise erklärt sich die p%-Regel ganz von selbst, als diejenige Schutzvorschrift, die bei ± f einen hinreichenden Intervallschutz für alle Einzelangaben garantiert.

Zur Begründung von (14) sei auf Abschnitt 3.2 verwiesen, wonach sich die Quaderspannweite, range, für einheitliche Schätzfehler $f \leq 100\%$ auch in positiver Tabelle gemäß $range = 2f \min X_j$ abschätzen lässt (mit(7) oder direkt mit (9)). Ein Quader Q ist demnach als Sicherungsquader immer zu verwerfen, wenn $2f * X_j / Pivot < q$ ausfällt, oder wenn $X_j / Pivot < q/(2f)$ für ein $X_j \in Q$ erfüllt ist (der Index j bezieht sich nur auf die Einzelwertgliederung). Die größte Einzelangabe X_1 als Pivot erzwingt also immer eine Sperrung ihres Veröffentlichungsaggregats X, wenn für den Quaderwert $X_j = X_3 \equiv Rest = X - X_1 - X_2$ eines ganz in den Mikrodaten liegenden Sicherungsquaders Q von X_1 gilt

$$(X - X_1 - X_2) / X_1 < q / (2f).$$

Dies ist genau die p%-Regel (0.4) zum Parameter $p = q / (2f)$. Der Quader Q zum Pivot X_1 würde allerdings aufgrund der Annahmebedingung $range / X_1 > q$ selbst bei Gültigkeit des Gleichheitszeichens in obiger Ungleichung noch verworfen. Die hinsichtlich der p%-Regel daraus resultierenden „Übersperrungen“ sind aber zu vernachlässigen. (Sie können im Übrigen durch Wahl eines infinitesimal größeren q, so dass in der Annahmebedingung auch Gleichheit gilt, vermieden werden.)

4.2.3.2 Relative Schätzfehler größer als 100%

Wenn der relative Schätzfehler f größer als Eins angesetzt werden kann, so ist die untere Schätzintervallgrenze in positiven Tabellen immer Null; also gilt für alle Quaderwerte $X_j \in Q : X_{o,j} - X_j = X_j f$ und $X_j - X_{u,j} = X_j$. Daraus ergibt sich für die Quaderauswahl die relevante Spannweite

$$range = \min [f \min X_i, \min X'_j] + \min [\min X_i, f \min X'_j] ; \quad i, j \in \{1, 2, 3\}$$

Man sieht daraus, dass die Quaderauswahl bei großen relativen Schätzfehlern, d.h. wenn $f \min X_i \geq \min X'_j$ und $f \min X'_j \geq \min X_i$ gilt, genau so erfolgt, als wäre nur die Vorinformation „positive Tabelle“ zu berücksichtigen und nicht auch eine relative Schätzfehlerspannweite. In diesem Falle ist nämlich $range = \min X'_j + \min X_i$, die Spannweite eines Quaders in positiver Tabelle. Nur, wenn einer der mit dem relativen Schätzfehler f multiplizierten Minimalwerte einer Quaderteilgesamtheit (der gerade indizierten oder der ungerade indizierten) kleiner als

der jeweils andere Minimalwert ist, geht der relative Schätzfehler zwingend in die Quaderauswahlbedingung range / Pivot > q ein:

Sei $f\,minX_i < minX'_j$, dann ist wegen $f > 1$ $minX_i < f\,minX'_j$, und aus obiger Spannweiteformel folgt range $= f\,minX_i + minX_i = (f+1)\,minX_i$. Zwar hat man in diese Beziehung zunächst nur den Minimalwert der gerade indizierten Quaderteilgesamtheit einzutragen, da aber wegen $f\,minX_i < minX'_j$ der gerade indizierte Minimalwert kleiner als jeder ungerade indizierte ist, ist er zugleich auch der kleinste Wert des betrachteten Quaders Q, so dass gilt

$$range = (1+f)\,minX_i\,, \qquad X_i \in Q$$

Im umgekehrten Fall $f\,minX'_j < minX_i$ erhält man das gleiche Ergebnis. Ein Quader Q zum Pivot X_p ist als Schutzquader geeignet, wenn im Falle $f > 1$ und $f\,minX_i < minX'_j$ oder $f\,minX'_j < minX_i$ gilt $(1+f)\,minX_i > q\,X_p$, $X_i \in Q$; er ist dagegen als Sicherungsquader zu verwerfen, wenn für ein $X_j \in Q$ gilt $X_j \le p\,X_p$, insbesondere auch für das Pivot $X_p = X_1$ und seinen Nachbarwert $X_j = X - X_1 - X_2 =$ Rest, wobei

$$p = q / (1+f)\,, \quad 1 < f \le minX^* / minX \qquad (15)$$

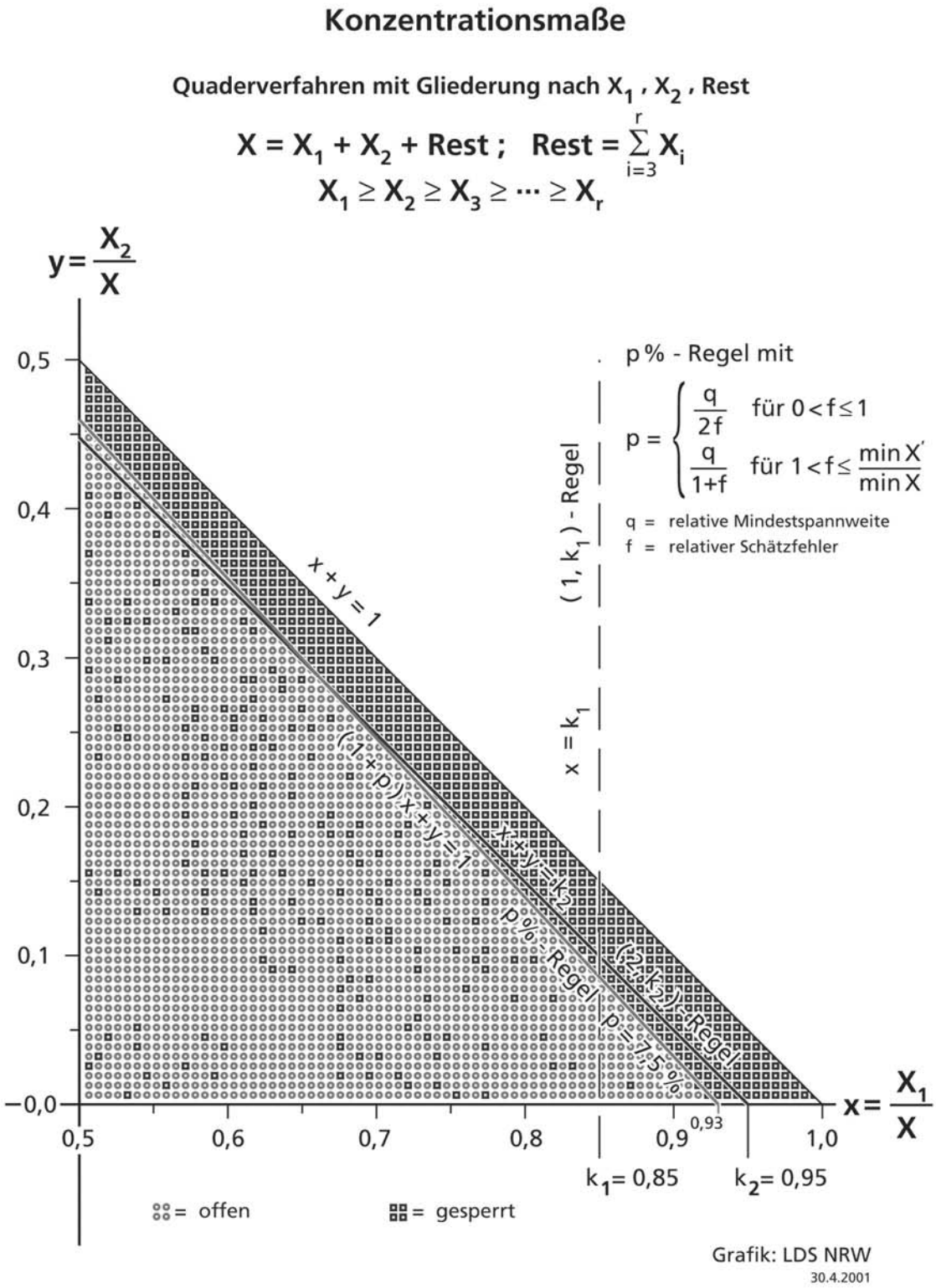
Während (14) bei $f \le 1$ immer angewendet werden kann, gilt (15) nur sehr bedingt! (15) ist anzuwenden, wenn $f\,minX_i < minX'_j$ oder $f\,minX'_j < minX_i$ gilt. Im Falle $f\,minX_i = minX'_j$ oder $f\,minX'_j = minX_i$ kann die Spannweite gemäß (5) für positive Tabellen berechnet werden, aber auch nach range $= (1 + f)\,min\,X$, $X \in Q$. D.h. die Bedingung für die Anwendbarkeit von (15) lautet $1 < f \le minX^* / minX$, wo $minX^*$ den jeweils größeren, minX den jeweils kleineren der Minimalwerte der beiden Quaderteilgesamtheiten bezeichnet. Das Auswahlkriterium ist dem gemäß von der Struktur des gerade betrachteten Quaders selbst abhängig.

Beispiel:

Gegeben sei eine Tabelle wie in Abbildung 4.1 mit einer zusätzlichen Zeile für von Null verschiedene Reste mit unterstelltem relativem Schätzfehler $f = 11$ für alle Tabellenwerte und vorgegebener relativer Mindestspannweite $q = 1{,}2$ für die Quaderauswahl. Der p%-Wert beträgt demnach $p = q/(1+ f) = 1{,}2/(1+ 11) = 0.1 = 10\%$. Ein Schutzquader des Pivots 1000, der die Veröffentlichungstabelle nicht trifft, wird über die Zeilensummenwerte in der rechten Randspalte und über die Restwerte geführt. Wenn der zu diesem Pivot gehörige Rest = 101 beträgt, so ist der Quader wegen Rest/Pivot = 101/1000 = 0.101 > p = 0,100 als Sicherungsquader für 1000 nach der p%-Regel nicht abzulehnen, das Verhältnis der beiden Minimalwerte $minX^* / minX = (Rest/Pivot)^{-1} = 9{,}9 < f = 11$ verbietet aber die Anwendung dieser Regel. Stattdessen ist die relative Quaderspannweite für positive Tabellen ohne Berücksichtigung von Schätzfehlern zu verwenden; sie beträgt $(minX^* + minX) / Pivot =$

$(1000 + 101) / 1000 = 1{,}101 < q = 1{,}2$. Der Quader ist also als Sicherungsquader abzulehnen. Die Sicherung muss mit einem über die Spaltensummen geführten Quader, der die unterste Tabellenzeile, die Veröffentlichungstabelle, trifft, vorgenommen werden.

Abb.4.2



Diese Beschränkung der Auswahlbedingung im Falle f > 1 auf ganz spezielle Quader hat ihre Ursache in der Positivität der Tabelle: Wenn der relative Schätzfehler f unbegrenzt zunimmt, geht der Auswahlparameter

$p = q / (1+f)$ gegen Null, womit jeder Quader als Sicherungsquader in Betracht käme. Das kann aber nicht richtig sein, weil bei unbegrenzt zunehmendem f immer noch die Vorinformation, „es liegt eine positive Tabelle vor“, zu berücksichtigen und damit $rang = \min X_i + \min X'_j$ anzuwenden ist. Im Falle $f > 1$ gibt es eben zwei Auswahlregeln[9].

Obige Graphik, Abbildung 4.2, gibt einen Überblick über die Verteilung der Primär- und Sekundärsperrungen in der Veröffentlichungstabelle. Sie zeigt nur die linke Hälfte des Konzentrationsmaß-Diagramms von Abschnitt 0.3, um so eine bessere Auflösung im Bereich der Primärsperrungen zu erreichen. Die kleinen grauen Kringel stellen offene Werte, die dunklen Vierecke Primär- und Sekundärsperrungen dar. Im Falle $f \leq 1$ trennt die Gerade $(1+p)\, x + y = 1$ genau die nach der p%-Regel oder durch die Quadersicherung mittels (14) gesetzten Primärsperrungen von den noch offenen oder sekundär gesperrten Werten der Veröffentlichungstabelle

Bei größeren relativen Schätzfehlern ist die „p%-Gerade“ keine Trennlinie mehr, sie wird gewissermaßen durchlässig für primär geheime Werte – soweit man die Unterscheidung zwischen Primär- und Sekundärsperrungen in der Veröffentlichungstabelle unter diesen Umständen überhaupt noch akzeptieren kann -, von denen mit zunehmendem f immer mehr in den Bereich unterhalb der p%-Geraden einwandern; oberhalb der p%-Geraden findet man nach wie vor nur Primärsperrungen. Das liegt daran, dass der mit (15), der unteren Formel in der geschweiften Klammer der Darstellung, berechnete p%-Wert als Grenzwert nur noch bedingt Gültigkeit hat und mit wachsendem f immer häufiger durch die Abschätzung mit der relativen Mindestspannweite für positive Tabellen ohne Schätzintervallangaben zu ersetzen ist.

Kann man relative Schätzfehler annehmen, die größer als 100% sind, so wird die p%-Regel mit der Anwendung des Quaderverfahrens auf die durch das Einzelmaterial erweiterte Tabelle also nur noch bedingt realisiert. Wenn der relative Schätzfehler schließlich als sehr groß gegenüber 100% anzunehmen ist, bleibt nur noch der Intervallschutz, der sich aus der Positivität der Tabelle ergibt. Ein Dominanzschutz ist im Falle sehr großer Schätzfehler wegen der zu ungenauen Abschätzbarkeit der Einzelangaben aber auch nicht erforderlich, weil mit den großen Schätzintervallen keine ausreichenden Kenntnisse über den Anteil vorhanden sind, den die Einzelwerte an ihren Aggregaten haben. Das Verfahren der Quadersicherung in durch das Einzeldatenmaterial erweiterten Tabellen ist also allgemeingültiger als die p%-Regel mit anschließender Quadersicherung, die es andererseits aber bei zu unterstellenden relativen Schätzfehlern von bis zu 100% auf eindrucksvolle Weise bestätigt.

Dem gemäß sollte auch das zu unterstellende Vorwissen über die Tabellenwerte und deren Dominanz mit Hilfe von Schätzintervallen eingetragen werden und nicht über von Dominanzmaßen erzwungene besonders große relative Mindestspannweiten, die dann immer wesentlich größer als Eins gewählt werden müssten (vergleiche „Wah-

---

[9] Durch Umformen von $(\min X'_j + \min X_i) / Pivot > q$ in $\min X_i / Pivot > q / (1 + \min X'_j / \min X_i)$ kann man zwar beide Auswahlkriterien über den Parameter $p_{total} = q / (1 + f_{min})$ mit $f_{min} = \min(f > 1, \min X'_j / \min X_i > 1)$ zu einem vereinigen, p_{total} ist darin aber kein vorgebbarer Parameter mehr, sondern hängt von der Struktur des jeweiligen Quaders ab.

rung der Geheimhaltung sensibler Daten in mehrdimensionalen Tabellen mit dem Quaderverfahren“ in „Statistische Analysen und Studien NRW“ ,3/2000). Die Vorinformation "Schätzintervalle" kann mit den neueren EDV-Programmen zum Quaderverfahren wie GHQUAR.45, GHMITER.22 oder QUIT ohne weiteres verarbeitet werden.

Mit der oben vorgestellten Möglichkeit der weitgehenden Einbeziehung der primären Geheimhaltung in das Quaderverfahren durch Intervallschutz von Einzeldaten besteht somit eine echte Alternative zu jedwedem Dominanzschutz in der primären wie auch in der sekundären Geheimhaltung. Intervallschutz von Einzeldaten ist eben immer auch Dominanzschutz der Einzelangaben in ihren Aggregaten; das Umgekehrte gilt nicht (vgl. 4.2.3.1, Text unter (14)). Dieses Verfahren bezieht insbesondere auch die unter 0.2.3 aufgeführten (p;q)-Regeln mit ein, weil – wie dort bereits erwähnt – diese Regeln unmittelbar in p%-Regeln überführt bzw. umgerechnet werden können. Dass der Mikrodaten-Intervallschutz auch für Tabellen mit nicht ausschließlich positiven Werten – für die bisher gar keine Dominanz definiert ist - einen ausreichenden Schutz gegen zu genaues Rückrechnen von Einzelwerten mit Hilfe von Veröffentlichungswerten bietet, versteht sich einfach aus dem allgemeinen Intervallschutz-Konzept der Quadersicherung.

4.3 Veröffentlichung von Schutzintervallen

Im Falle der in der amtlichen Statistik häufig auftretenden sehr fein gegliederten Tabellen werden aufgrund der mit der Feinheit der Tabellengliederung einhergehenden „dünnen“ Besetzung sehr viele Sperrungen eingetragen, die in der Veröffentlichungstabelle dann beispielsweise als Sternchen erscheinen. Eine derart von Schutzsternchen perforierte Veröffentlichungstabelle mindert die Akzeptanz solcher Statistiktabellierungen, wobei man allerdings meist nicht bereit ist, die Tabellengliederung zwecks Einsparung von Sperreintragungen zu vergröbern.

Einer inzwischen häufiger geäußerten Anregung folgend, kann man aber anstelle der Schutzsternchen für jeden geheimen Tabellenwert seine Schutzintervallgrenzen eintragen, die jeder Tabellennutzer mit Hilfe von linearen Optimierungsprogrammen selbst errechnen könnte, wenn diese nicht so aufwändig in der Handhabung wären. Um den Tabellennutzer zu unterstützen, hat das Statistische Bundesamt ein EDV-Programm in Auftrag gegeben, das für gesicherte Tabellen solche Schutzintervallgrenzen auf der Basis der linearen Optimierung berechnet.

Lineare Optimierungen sind i.A. sehr rechenzeitaufwändig. Obwohl der Aufwand bei einmaliger Berechnung von Schutzintervallen keineswegs vergleichbar ist mit der sehr oft zu wiederholenden Schutzintervallberechnung bei der Auswahl des günstigsten Sperrmusters (vgl. 3.1.1), ist er doch weit höher als beim Quaderverfahren, wo solche Schutzintervalle bereits nach Abarbeitung der Tabelle vorliegen: Wie in Abschnitt 3 beschrieben, wird für die Quaderauswahl die Quaderspannweite berechnet. Dabei werden immer auch die Schutzintervallgrenzen jedes Quaderwertes bestimmt. Durch diese Intervallgrenzen werden für jeden Wert eines vollständigen Quaders (vgl. 3.2.3.3) Schutzintervalle beschrieben, die ein externer Tabellennutzer durch den Einsatz von linearen Optimierungsverfahren nicht weiter eingrenzen kann - das gilt auch für das o.g. Optimierungsverfahren zur Schutzintervallberechnung -. Andererseits erfolgt die Quaderauswahl gerade so, dass die Schutzintervalllänge, die Quaderspannweite, für den Schutz der geheimen Werte gegen zu genaues Rückrechnen ausreicht. Die Schutzintervalle der

Werte vollständiger Quader brauchen also nicht noch einmal gesondert berechnet, sondern können direkt, wie schon unter 3.2.3.3 beschrieben, unter Berücksichtigung der Intervall-Ausgabe-Regel veröffentlicht werden.

Der Tabellennutzer erhält dadurch eine zusätzliche Information über den Bereich, den ein gesperrter Tabellenwert überdecken kann, was die Datensicherung akzeptabler macht, ohne dabei den erforderlichen Datenschutz in Frage zu stellen. Dieser Komfort kann allerdings nur für vollständige Quader geboten werden und nicht, wenn bei dem betreffenden Quader noch mit Untertabellenabgleich oder mit Abgleich anderer Veröffentlichungstabellen gearbeitet werden muss. Sollen auch Schutzintervalle von mehreren Tabellen gemeinsamen Quaderwerten ausgedruckt werden, wird man wohl nicht um o.g. Optimierungsverfahren herumkommen. Dabei wird allerdings erst die Praxis erweisen müssen, ob die bisher durch Abgleich bearbeiteten überlappenden Tabellen tatsächlich als Einheit mit obigem Optimierungsverfahren bearbeitet werden können (Rechenzeit, akzeptable Nutzeroberfläche des Optimierungsverfahrens).

5. Justierung der Verteilung von Sekundärsperrungen

Bei gegebener Tabellenstruktur und gegebener Verteilung primär geheimer Werte wird die Verteilung der Sekundärsperrungen auf die noch offenen Tabellenwerte durch die Quaderauswahlregeln des allgemeinen Sicherungskonzepts für n-dimensionale Tabellen (Punkt 2.1) in Verbindung mit den Intervallschutzregelungen (Punkt 3.1) oder bei Berücksichtigung vom Nutzer vorgebbarer Schätzintervalle (Punkt 3.2) vollständig festgelegt. Bei einigen Anwendungen besteht aber von Seiten des Statistikers ein fachlich durchaus begründetes Interesse, die durch obige Regelungen bestimmte Auswahl von Sekundärsperrungen zu verändern, um sie an die fachlichen Gegebenheiten anzupassen. So können manche noch offenen Tabellenwerte zum Schutze anderer nicht gesperrt werden, weil sie der Öffentlichkeit ohnehin bekannt sind, als gesperrte Werte also keinen Schutz bieten; oder es sollen gewisse regionale Einheiten zur Entlastung anderer besonders bevorzugt gesperrt werden usw..

Um die Auswahl der Sicherungsquader und damit das Muster der Sekundärsperrungen weitgehend den fachlichen Gegebenheiten anzupassen, wird man die Eingabedaten geeignet modifizieren bzw. gewichten, was man als Uminterpretation der in den Daten enthaltenen Information deuten kann. Dabei wird die zu minimierende Summe zu sperrender Werte des jeweils zur Auswahl stehenden Quaders, der zu minimierende Informationsverlust für einen geschützten geheimen Quaderwert (Pivot), entsprechend verändert. Die mit der ersten Priorität belegte Minimierung der Anzahl der Sekundärsperrungen steht nicht zur Disposition, obgleich sie von der Quaderwertesummenminimierung und deren Modifikation nicht ganz unberührt bleiben wird.

Dieser Abschnitt gibt einen Überblick über die enorme Vielfalt der Modifikationsmöglichkeiten des Quaderauswahlverfahrens, wobei ein gewisser Bezug zu spezifischen Eigenschaften der EDV-Programme GHQUAR, GHMITER oder QUIT unvermeidbar sind, weil dabei auch die Speicherplatzorganisation eine wichtige Rolle spielt.

5.1 Vorübergehende Veränderung der Eingabedaten

5.1.1 Vorübergehende Veränderung der Anzahl der Nachweisungsfälle

Die Anzahl der Nachweisungsfälle ist nur in Kontingenztabellen als zu minimierende Größe (Zielfunktion) zu behandeln. In so genannten Wertetabellen, die Nachweisungsfälle und die von diesen berichteten Werte ausweisen, ist die zu sperrende Wertesumme die zu minimierende Zielfunktion. Trotzdem hat die Fallzahl gemäß Punkt 2.1 auch in Wertetabellen Einfluss auf die Auswahl sekundär zu sperrender Werte. Nach Punkt 2.1 (siehe dazu insbesondere 2.1.2) ist das Quaderverfahren so angelegt, dass nach Möglichkeit keine Tabellenfelder mit nur einem Berichtenden als Sicherungspositionen (Quaderwerte) ausgewählt werden. Bei ungünstiger Verteilung der Einzelangaben, wo eine sonst unerwünscht hohe Anzahl von Sicherungssperrungen unter Umständen auch in die Randsummen zu erwarten wäre, kommen auch Einzelangaben als Sicherungspartner in Betracht. Dann muss aber nach 2.1.2 ein weiterer Sicherungsquader, der die Einzelangaben des ersten nicht enthält, ausgewählt werden.

Einzelangaben stellen ein erhöhtes Sicherheitsrisiko dar, dem mit der Doppelquadersicherung begegnet wird. Wenn aber der für die Sicherung sensibler Daten Verantwortliche keine Notwendigkeit sieht, die Einzelangaben besonders zu schützen (weil er zum Beispiel in Veröffentlichungstabellen nicht nur die geheimen Werte selbst, sondern auch deren Fallzahlen sperrt) kann er die Doppelquadersicherung durch temporäres Verändern der Fallzahlen ausschalten. Beim Einsatz eines EDV-Programms genügt es zum Beispiel - nur für die Dauer der Bearbeitung mit dem Quaderverfahren - alle Fallzahlen, die nur einen Berichtenden anzeigen, durch die Fallzahl = 2 zu ersetzen oder zu allen von 0 verschiedenen Fallzahlen die Zahl 1 zu addieren, um zu erreichen, dass keine Einzelfälle mehr berücksichtigt werden. Dann sind keine Doppelquadersicherungen mehr erforderlich und ehemalige Einzelangaben dürfen uneingeschränkt als Sicherungspartner eingesetzt werden. Diese Maßnahme reduziert dann die Anzahl der Sekundärsperrungen gerade bei schwach besetzten Tabellen beträchtlich.

5.1.2 Vorübergehende Veränderung der berichteten Tabellenwerte

5.1.2.1 Behandlung von Tabellen mit positiven und negativen Werten

Alle bisher realisierten EDV-Verfahren zur Wahrung der Geheimhaltung, die auf dem Quaderverfahren basieren, sind für so genannte positive Tabellen konzipiert worden. Um mit diesen Verfahren auch Tabellen, die sowohl positive als auch negative Werte enthalten, bearbeiten zu können, geht man davon aus, dass kein Intervallschutz erforderlich ist (es liegen keine externen Schätzintervalle vor) und transformiert die Tabelle in eine positive Tabelle. Dies kann auf zweierlei Weise geschehen:

1. In Tabellen mit negativen Werten können - nur für die Bearbeitung mit dem Geheimhaltungsverfahren - die absoluten Beträge der Tabellenwerte anstelle der Werte selbst eingetragen werden. Um die Null als gleichberechtigten Sperrkandidaten zu verwenden, ersetzt man Nullwerte durch den von Null verschiedenen minimalen Wert der absoluten Beträge der Tabellenwerte. Dadurch wird erreicht, dass Nullwerte in beiden Quaderteilgesamtheiten, der gerade und der ungerade indizierten, auftreten können, ohne den betreffenden Quader als Sicherungsquader für einen zu schützenden geheimen Wert ausschließen zu müssen.

2. Bei der Behandlung von Tabellen mit negativen Werten kann auch zu allen Tabellenwerten die Summe aus dem absoluten Betrag des kleinsten Wertes und eines berechneten Minimalwertes addiert werden. Als berechneter Minimalwert dient bisher der 10^{8}-te Teil des maximalen Wertes der absoluten Beträge der Tabellenwerte als Erfahrungswert. (Der Faktor 10^{-8} entstammt der Umsatzsteuerstatistik NRW 1996, wo das Verhältnis aus minimalem und maximalem Tabellenwert größer als 10^{-8} ist). Das heißt, es erfolgt eine Werteverschiebung, die bewirkt, dass die in den Tabellenwerten enthaltene Information als Abstand vom kleinsten Tabellenwert (und nicht von der Null) gemessen wird. Nullwerte werden auch hier als zu allen anderen Werten völlig gleichberechtigt behandelt.

Beide Vorgehensweisen sind in den letzten Programmversionen von GHQUAR., GHMITER und QUIT programmintern realisiert und können optional gesteuert werden.

Wie bereits bemerkt, sind Tabellen, bei denen die Information über ihre Positivität bzw. ihre externen Schätzintervalle fehlt, leichter zu sichern als positive Tabellen. Es ist jedoch Vorsicht geboten. Man darf eine positive Tabelle, bei der also jedem Nutzer von vornherein bekannt ist, dass sie keine negativen Werte enthalten kann, niemals als „nicht ausschließlich positive“ deklarieren, weil durch den dann anzuwendenden wesentlich reduzierteren Schutz Geheimhaltungslücken aufträten. Dies hängt mit der unterschiedlichen Behandlung von Nullwerten zusammen und wird unter Punkt 5.1.2.3 eingehend erläutert.

5.1.2.2 Ersetzen von Tabellenwerten durch andere Werte

Um zu erreichen, dass spezielle Werte bei der Auswahl von Sperrpositionen besonders bevorzugt bzw. besonders benachteiligt werden, kann man die Eingabewerte durch beliebige andere Werte, z.B. sehr kleine bzw. sehr große von Null verschiedene Werte, ersetzen. Spezialfall: Ersetzen aller Tabellenwerte durch einen Einheitswert, wenn der Informationsverlust durch Sperren von Werten nicht an der Wertgröße gemessen werden soll, oder Ersetzen der Werte durch die Anzahl der Berichtenden, wenn nicht der Wert selbst, sondern die Anzahl der Meldenden für den Informationsverlust durch Sekundärsperrungen von Bedeutung ist. Diese Veränderung von Tabellenwerten ist im Allgemeinen nicht mit dem Intervallschutz zu kombinieren, weil die Spannweitenberechnung mit verfälschten Werten durchzuführen wäre, was zu falschen ranges und damit auch zu einer fehlerhaften Auswahl von Schutzquadern führen würde.

Die Justierung der Verteilung der Sekundärsperrungen durch Verfälschung der Eingabedaten ist bei Tabellen, die mit Intervallschutz gesichert werden sollen, also grundsätzlich abzulehnen!

5.1.2.3 Einführung von sperrbaren Nullen

Wie im Kapitel 3 unter Punkt 3.1, aber auch unter 3.2 (insbesondere unter 3.2.2) ausgeführt, kommen auch Tabellenfelder mit Wert Null als Sicherungsfelder für Sperrungen in Betracht. Dazu eignen sich allerdings nicht alle Nullwerte, sondern nur solche, bei denen davon ausgegangen werden kann, dass ihr Wert der Öffentlichkeit nicht bekannt ist; die anderen so genannten strukturellen Nullen sind als Sperrkandidaten auszuschließen. Um die beiden Arten von Nullwerten bei der Durchführung der sekundären Geheimhaltung voneinander zu unterscheiden, werden die sperrbaren Nullen mit einem durch die Steuerung des Quaderverfahrens festgelegten symbolischen Wert in die Eingabedaten eingebracht und bei der Spannweitenberechnung als Wert Null berücksichtigt. Durch die Einführung von sperrbaren Nullen wird erreicht, dass im Falle dünn besetzter Tabellen bei der Quaderauswahl (auch bei Intervall- und Dominanzschutz) weniger häufig auf Randsummenwerte ausgewichen werden muss.

Besonders effektiv ist die Freigabe von nicht strukturellen Nullwerten als Sperrkandidaten bei Tabellen, die sowohl positive als auch negative Werte ausweisen und die keinen Intervallschutz erfordern, weil in diesen Tabellen Nullwerte gleichzeitig sowohl in der gerade indizierten als auch in der ungerade indizierten Quaderteilgesamtheit auftreten können, ohne diesen Quader als Schutzquader ausschließen zu müssen. Hier macht sich der Informationsverlust bei nicht positiven Tabellen gegenüber Tabellen mit Vorinformation - wie z.B. Positivität der Tabelle oder vom Nutzer angebbare Schätzintervalle - konkret bemerkbar:

Betrachtet man beispielsweise eine eindimensionale Tabelle, in der einem primär geheimen Nullwert ein anderer geheimer Nullwert als Quaderwert zugeordnet ist, so ist dieser aus zwei Nullen bestehende Quader in einer positiven Tabelle als Schutzquader ungeeignet. In einer nicht positiven Tabelle ohne die Vorinformation von Schätzintervallen aber genügt dem Tabellennutzer das Wissen, dass die Werte beider Tabellenfelder in ihrer Summe Null ergeben, nicht, um daraus jeden der beiden Einzelwerte zu schätzen; sie könnten beide nämlich Null sein, sie könnten sich aber auch aufgrund unterschiedlichen Vorzeichens gegenseitig kompensieren, und das bei beliebigen Wertebeträgen. In einer Tabelle ohne zusätzliche Information (Positivität, Nutzer-Schätzintervalle) können alle Quaderwerte aus nicht strukturellen Nullen bestehen, ohne diesen Quader als Sicherungsquader ausschließen zu müssen.

Zusammenfassend lässt sich also sagen, dass die Einführung von sperrbaren Nullen dünn besetzte Tabellen bis zu einem gewissen durch die Teilquaderstruktur bestimmten Grade mit Sperrkandidaten auffüllen kann, was zu einer Reduzierung der Randsummensperrungen beiträgt (Repsilber, Luxemburg 1994).

5.1.2.4 Weglassen von Tabellenwerten bzw. ganzen Tabellenteilen

Um zu erreichen, dass vorgegebene Tabellenwerte niemals gesperrt werden, kann man sie und ihre Fallzahlen weglassen, d.h. entsprechende Tabellenfelder durch strukturelle Nullen ersetzen. Dabei dürfen allerdings keine Strukturbrüche in Bezug auf die Summen- und Zwischensummenstruktur der Tabelle auftreten. Es ist z.B. auszuschließen, dass eine Summe über leere Tabellenfelder einen von Null verschiedenen Tabellensummenwert ergibt; umgekehrt darf die Summe aus lauter positiven Werten nicht zu einem Summenwert Null führen (eine exakte Summenüberprüfung erfolgt aber in den EDV-Programmen des LDS NRW nicht).

Unter dieser Voraussetzung kann man auch einen Tabellenwert weglassen, zu dem derselbe Berichtende beiträgt wie zu einem anderen ebenfalls in der Tabelle vorhandenen Wert, und dann die Tabelle sichern. Anschließend fügt man den Wert wieder ein und lässt statt dessen den anderen Wert weg und sichert erneut. Auf diese Weise können Tabellen mit Werten, die ganz oder teilweise auf dieselben Berichtenden zurückgehen, wie Tabellen mit lauter verschiedenen Berichtenden in allen Feldern behandelt werden. Dieses Vorgehen ist beispielsweise bei Doppelquadersicherung angezeigt (vgl. 2.1.2); dabei müssen die Einzelangaben des einen Quaders ohne Pivot von anderen Berichtenden stammen als die des zweiten Quaders.

Zur Vermeidung besonders vieler Sperrungen in schwach besetzten Tabellen können ganze Tabellenteile, die bezüglich einer Gliederung zur selben Summe beitragen, weggelassen werden. Man kann beispielsweise auf Gemeindeebene die Tabellenfelder (Datensätze der Eingabedatei) weglassen, die zum selben Kreis beitragen; der Kreis (ohne sein Hinterland) wird dann wie eine kreisfreie Stadt behandelt; ihre „Gemeinden“ dürfen dann aber niemals veröffentlicht werden.

5.2 Programminterne Justierung

Bei allen Maßnahmen zur Justierung der Verteilung sekundär geheimer Werte ist zu beachten, dass bei Spannweiteberechnungen zur Realisation des Intervallschutzes für die primär geheimen Werte immer nur die ursprünglichen, unveränderten Tabellenwerte benutzt werden dürfen. Das bedeutet, dass die Information über die Wertbeträge auch nach einer Tabellenwertmodifikation für das Geheimhaltungsverfahren weiterhin zur Verfügung stehen muss.

5.2.1 Wertestaffelung und Randsummengewichtung

In den bisher realisierten Geheimhaltungsprogrammen wird die für die Spannweitenberechnung benötigte Wertinformation durch eine logarithmische Klassierung der Tabellenwerte mit hinreichender Genauigkeit konserviert: Durch mehrfache Verschiebung dieser endlichen Klassengesamtheit um deren Spannweite entlang der Zahlen-Achse entsteht eine disjunkte hierarchische Staffelung von Klassenwerten. Liegen beispielsweise die Klassenwerte im Intervall (0; 100 000], so ergibt sich die Staffelung durch Addition eines ganzzahligen Vielfachen von 100 000 zu jedem Klassenwert:

----, (- 200 000; - 100 000], (- 100 000; 0], (0; 100 000],----

Alle Klassenwerte einer betrachteten Hierarchiestufe (Staffel) sind stets größer als alle Klassenwerte der niedrigeren Hierarchiestufen. Mit dieser Hierarchiestaffelung hat man ein Mittel zur diskreten Gewichtung der Tabellenwerte. Außerdem können auch gewisse Wertattribute, wie z.B. die Eigenschaft, ein primär geheimer Wert oder eine Einzelangabe zu sein, durch die Staffelzugehörigkeit angezeigt werden, so dass der Zugriff auf die solchermaßen gestaffelten Werte immer auch den gleichzeitigen Zugriff auf die Attributstabellen beinhaltet, wodurch Zugriffszeit gespart wird.

Dazu werden die Tabellenklassenwerte in die jeweiligen, ihren vorgesehenen Gewichtungen bzw. ihren Attributen entsprechenden Hierarchiestufen eingegliedert. Die für die range-Berechnung erforderliche Wertinformation liegt weiterhin in der Klassierung innerhalb der Hierarchiestufen. Für das Summenkriterium wird aber in erster Linie die Zugehörigkeit des betreffenden (Klassen-)Wertes zu seiner Hierarchiestufe wirksam und erst in zweiter Linie seine Position innerhalb der Hierarchiestufe.

Darüber hinaus können in allen EDV-Programmen des LDS NRW Randsummen durch Setzen von Randschranken höher oder niedriger als Tabellenwerte im Inneren der Untertabellen gewichtet werden, um zu erreichen, dass das Programm mit Randschranken belegte Summenwerte nur dann sperrt, wenn keine anderen Möglichkeiten der Quadersicherung bestehen; oder, bei niedrigerer Gewichtung, die betreffenden Randsummen besonders bevorzugt sperrt.

Die Beträge der Schrankenwerte sind von der Anzahl an der Randsumme beteiligter Gliederungsmerkmale abhängig, d.h. dimensionsabhängig, und einheitlich für alle Gliederungskriterien. Für jedes Gliederungskriterium kann - unabhängig von den anderen - die Randschranke positiv oder negativ oder auch nicht gesetzt werden; eine kontinuierliche Randsummengewichtung ist zwar möglich aber nicht sinnvoll, weil bei unterschiedlich hoher kontinuierlicher Gewichtung die erprobte dimensionsabhängige Quaderauswahl gestört werden könnte.

5.2.2 Auszeichnung geheimer Werte

Um zu erreichen, dass das Summenkriterium zu Gunsten von Quadern mit möglichst vielen bereits gesperrten Tabellenwerten ausfällt, wobei hier die Summe aller Quaderwerte minimiert wird, kann man geheime Werte durch tabellendimensionsabhängige, betragsmäßig große negative Werte ersetzen. In diesem Falle vermeidet eine temporäre, nur für die Quadersummenberechnung vorgenommene Ersetzung der betreffenden Klassenwerte durch einen betragsmäßig großen einheitlichen negativen Wert die unterschiedliche Bewertung aufgrund der unterschiedlichen Klassenpositionen innerhalb der Hierarchiestufe.

Das entspricht genau der Zielsetzung, mit höchster Priorität die Anzahl der Sekundärsperrungen so klein wie möglich zu halten, d.h. Sicherungsquader auszuwählen, die möglichst viele bereits gesperrte Werte enthalten, wobei die Größe des geheimen Wertes keine Rolle spielen darf. Die Wertgröße wird erst für die range-Berechnung wichtig; dazu steht die benötigte Information in Form des entsprechenden Klassenwertes aus einer Klasse von geheimen Werten zur Verfügung.

Bei der Bemessung des den geheimen Werten in der Quadersumme zuzuschreibenden Alternativ-Wertes muss man berücksichtigen, dass ein Quader mit 2^n-1 offenen sehr kleinen positiven Werten als Sicherungspartner des zu schützenden geheimen Pivots eine größere Quadersumme ergeben soll als ein Quader mit 2^n-2 sehr großen positiven Partnerwerten und nur einem geheimen Partnerwert.

Die Staffel der offenen Tabellenwerte sei durch das Intervall $(K_{min}\ ;\ K_{max}]$ gegeben, wo K_{max} den größten und $K_{min} \geq 0$ den kleinsten Klassenwert der noch offenen Tabellenwerte bezeichnet. K_g sei der einem geheimen Tabellenwert in der Quadersumme zuzuordnende Wert, so ist obige Forderung erfüllt, wenn die Ungleichung

$$(2^n\text{-}1)*K_{min} > (2^n\text{-}2)*K_{max} + K_g$$

gilt und dies trifft jedenfalls zu, wenn

$$K_{g1} = -\ (2^n\text{-}1)*K_{max}$$

für K_g gesetzt wird. Im Falle einer eindimensionalen Tabelle mit verschwindendem kleinsten Klassenwert könnte für K_g nach obiger Ungleichung eine beliebige negative Zahl gewählt werden; die Ersetzung von K_g durch K_{g1} legt diese negative Zahl auf $-K_{max}$ fest.

Um bei der Sicherung bevorzugt geheime Tabellenwerte zu verwenden, die von mehr als einem Berichtenden gemeldet wurden, hat sich für diese ein zusätzlicher Faktor von 1,1 bewährt:

$$K_{g2} = -1{,}1*(2^n-1)*K_{max}$$

Einzelangaben werden also mit dem Wert K_{g1} und andere geheime Werte mit K_{g2} in die Quadersumme eingetragen. Dadurch wird die Einzelquadersicherung gegenüber einer Doppelquadersicherung bevorzugt.

Ganz ausgeschlossen ist die Doppelquadersicherung bei diesem Vorgehen selbst dann nicht, wenn auch ein Einzelquader zum Schutze eines geheimen Wertes zur Verfügung gestanden hätte (der eigentlich nach den Regelungen vom Abschnitt 2.1.1 hätte bevorzugt werden müssen). Die mit obigem Faktor herbeigeführte Bevorzugung von Einzelquadern gegenüber einer Doppelquadersicherung genügt aber bei praktischen Anwendungen und führt im statistischen Mittel erfahrungsgemäß sogar zu besonders wenigen Sekundärsperrungen.

Es sei nochmals darauf hingewiesen, dass bei obigen Wertemanipulationen nur der Punkt 5.1.2.2 eine kontinuierliche Gewichtung ermöglicht, und das auch nur bei Verzicht auf Intervallschutz, alle anderen unter 5.1 und 5.2 angeführten Justierungsmaßnahmen sind diskreter Art und gewähren auch Intervallschutz.

Um die Tabellenwerte auch bei Gewährleistung von Intervallschutz kontinuierlich gewichten zu können, müssen die ursprünglichen unveränderten Tabellenwerte auch nach der vorgenommenen Gewichtung weiterhin mitgeführt werden. Dies ließ sich am einfachsten durch Einführung komplexer Tabellenwerte lösen, weil dabei die ursprüngliche Datensatzstruktur der n-dimensionalen Tabelle, n Gliederungsmerkmale, Fallzahl, (komplexer) Wert und Sperrschlüssel beibehalten werden kann und alle in dem Tabellenwert verschlüsselten Einzelwerte wie klassierter Tabellenwert, standardisiertes Gewicht dieses Tabellenwertes mit einem einzigen Daten-Zugriff erfasst werden können (vgl. Übersicht zur Struktur des Wertfeldes der Gesamttabelle im Hauptspeicher am Ende von 3.2.3.2).

In einem ersten Ansatz wurde der klassierte Tabellenwert in den Realteil, sein Gewichtswert in den Imaginärteil der komplexen Zahl eingetragen. Die so erweiterte EDV-Programm-Version, GHQUAR, Version 4 bot damit das erste Mal die Möglichkeit einer freien kontinuierlichen (und sogar auch negativen) Gewichtung, ohne dabei auf den originalwertebezogenen Intervallschutz verzichten zu müssen. Ab GHQUAR.45 werden in allen Folgeprodukten doppeltgenaue komplexe Zahlen als Tabellenwertvariable genutzt, die im Realteil den klassierten Wert, sein (ausschließlich positives) Gewicht und den größten Einzelwert aufnehmen. Der Imaginärteil speichert die obere und die untere Schätzfehlergrenze.

Alle folgenden Betrachtungen beziehen sich auf GHQUAR-Versionen mit komplexer Wertevariabler einfacher Genauigkeit. Insbesondere die in diesen Programmversionen erlaubte negative Gewichtung von Tabellenwerten gestattet eine direkte Bearbeitung von mit Dummy-Werten aufgefüllten aufgestockten vollständigen Tabellen, was bei der Diskussion von Sperrmustern vollständiger Tabellen ausgiebig genutzt wird.

Die Liste der verfügbaren Justierungsmaßnahmen ist nun noch um einen dritten Punkt zu ergänzen.

5.3 Justierung durch externe Gewichtung

Alle neueren, im LDS NRW entwickelten EDV-Programme zum Quaderverfahren, GHQUAR ab Version 4, GHMITER und QUIT, sehen im Satzformat des Eingabebestandes ein zusätzliches numerisches Tabellenfeld für eine externe Gewichtung vor. In dieses Feld kann im Falle von GHQUAR ein beliebiger reeller Zahlenwert (zwischen -10^{50} und $+10^{50}$) eingetragen werden, GHMITER und QUIT akzeptieren nur positive reelle Werte. Mit diesen Werten werden dann die klassierten offenen Tabellenwerte bei der Berechnung der Quadersumme multipliziert, d.h. gewichtet.

Beispielsweise ist bei GHQUAR ein solchermaßen gewichteter offener Wert als Sicherungspartner in einem zweidimensionalen Quader attraktiver als ein primär geheimer Wert (einschließlich Einzelangaben), wenn sein „Gewicht“ kleiner als ein Schrankenwert von -330000 ist. Bei höherdimensionalen Tabellen erhöht sich der Betrag dieses Schrankenwertes dimensionsabhängig entsprechend der Zunahme der Quaderwerte, siehe 5.2.2.

Ein sehr großer Gewichtswert, z.B. 10^{50}, den man sowohl bei GHQUAR als auch bei GHMITER und QUIT in den Eingabebestand eintragen kann, bewahrt einen offenen Tabellenwert vor einer Sekundärsperrung weitgehend. Eine exakte Aussage, ab welcher Gewichtsgröße ein offener Tabellenwert mit Sicherheit offen bleibt, ist i. Allg. nicht zu machen, weil u. U. auch andere, ebenfalls hochgewichtete Werte mit dem betrachteten gewichteten Tabellenwert konkurrieren. Soll keine Gewichtung vorgenommen werden, ist das Gewicht = 1 in das entsprechende Datenfeld des Eingabe-Datensatzes einzutragen.

5.3.1 Vorgabe von Gewichtsfunktionen

Die große Vielgestaltigkeit der freien externen Gewichtung lässt sich dadurch überschaubarer machen und damit auch besser automatisieren, dass man für die vorzugebenden Gewichtszahlen funktionale Zusammenhänge mit den Tabellenwerten, der Anzahl von Berichtenden sowie den Tabellenfeldpositionen aufstellt. Hier seien nur einige Beispiele als Anregungen aufgeführt:

a) Nach dem Summenkriterium wird bei positiven Tabellen Gleichwertigkeit bei der Auswahl von zu sperrenden Tabellenwerten erreicht, wenn man für die Tabellenwerte den Reziprokwert des Logarithmus aus dem auf den Tabellenminimalwert bezogenen Tabellenwert als Gewichtsfunktion verwendet. Diese Art der externen Gewichtung ist dann anzuwenden, wenn der Informationsverlust durch Sperren von Tabellenwerten als von der Größe der Tabellenwerte unabhängig angenommen werden soll: Es kommt dem Anwender nur darauf an, dass ein Tabellenwert gesperrt werden muss und nicht darauf, wie groß der ist.

b) Soll die Anzahl der Berichtenden die Sperrpositionen mitbestimmen, wobei auch der Betrag des Tabellenwertes von Einfluss sein soll, so bietet sich als Gewichtsfunktion die Anzahl der Berichtenden an; soll aber die Anzahl der Berichtenden allein das Ausmaß des Informationsverlustes durch die Sekundärsper-

rungen beschreiben, so wird man die Anzahl der Berichtenden durch den entsprechenden Klassenwert des offenen Tabellenwertes dividieren und als Gewicht in das Eingabefeld des Datenbestandes eintragen. - Bei positiven Tabellen kann man – nach den Muster von a) - den Klassenwert durch den Logarithmus des jeweiligen auf den Tabellenminimalwert bezogenen Tabellenwertes ersetzen.

c) Als Beispiel für eine von der Position der Tabellenfelder abhängige Gewichtung ist die externe Randsummengewichtung aufzuführen, die hier - anders als bei interner Justierung nach Pt. 5.2.1 - unterschiedlich für verschiedene Gliederungskriterien oder auch für verschiedene Gliederungsmerkmalsgruppen vorgegeben werden kann. Ein anderes Beispiel für die tabellenfeldpositionsabhängige Gewichtung ist durch die Auswahl von gewissen Tabellenfeldern oder auch ganzen Tabellenteilen durch auf die Gliederungskriterien wirkende Auswahlkriterien gegeben; solche Tabellenteilgesamtheiten können dann einheitlich gewichtet oder auch mit Gewichten einer geeigneten Gewichtsfunktion z.B. gemäß a) und b) belegt werden.

Technische Anmerkung

Alle Gewichtungsmaßnahmen zur Justierung des Sperrmusters einer Statistiktabelle bewirken, dass die für die Auswahl von Sicherungsquadern wichtige Quaderwertesumme verändert wird. Dies geschieht durch die Veränderung der einzelnen Summanden mit Hilfe der Gewichtsfaktoren. Dabei muss man berücksichtigen, dass die gewichteten Klassenwerte nicht zu weit auseinander klaffen, weil sonst betragsmäßig besonders kleine Summanden mitunter gar nichts mehr zur Unterscheidung der Quadersummen beitragen. Das Summenkriterium ist dann in Bezug auf die durch die Gewichtung erhaltenen betragsmäßig kleinen Quadersummen-Werte außer Kraft gesetzt.

Durch die EDV-mäßige Realisierung des gewichteten Quaderverfahrens wird dies besonders deutlich: Erfolgt dabei die Summation über REAL*4-Werte, so ist das Ergebnis nur für sieben wesentliche Dezimalstellen richtig wiedergegeben. Wenn dann eine Gesamtheit von Quadern existiert, die alle denselben gewichteten Klassenwert 10^{50} gemeinsam haben und deren restlichen gewichteten Werte beispielsweise einen kleineren Betrag als 10 Milliarden haben, so sind alle Quader dieser Gesamtheit ununterscheidbar; ihre Quaderwertesumme ist einheitlich 10^{50}. Bei der Sicherung wird daher der erste „beste“ Quader dieser Gesamtheit ausgewählt, d.h. die Quaderauswahl ist in solchen Fällen eine eher zufällige. Eine werteunabhängige Quaderauswahl kann sachlich begründet sein. Im Allgemeinen wird man aber eine wertegesteuerte, weitgehend eindeutige Quaderauswahl bevorzugen. Daher empfiehlt sich meistens eine moderate Vergabe von Gewichten, mit Beträgen etwa im Bereich von 1 bis 100, weil dabei Rundungsverfahren noch keinen wesentlichen Einfluss haben.

5.3.2 Externe Gewichtung zur Bearbeitung von Zeitreihentabellen

Eine für die amtliche Statistik besonders interessante Anwendung der externen Gewichtung ist die Sicherung von zeitperiodischen Statistiktabellen gegen zu genaue Rückrechnung sensibler Tabellenwerte (z.B. monatlich zu veröffentlichende Statistiken). Es ist hier anzumerken, dass das „Ordnungskriterium“ Zeit in Zeitreihentabellen hinsichtlich der sekundären Geheimhaltung nicht als zusätzliche Dimension gesehen werden darf, denn mit dem

Tabellenparameter Zeit ist keine Summenbeziehung verknüpft. Umgekehrt werden im Folgenden alle Ordnungskriterien, die – wie die Zeit – keine Summenbeziehungen unterhalten, als zeitliche bezeichnet und die danach geordneten gleichartig strukturierte Tabellen als Zeitreihen. Alle diese Zeitreihen können mit nachstehendem Verfahren bearbeitet werden.

Das Problem der sekundären Geheimhaltung in zeitperiodischen Tabellen besteht darin, dass gesperrte Werte in der aktuellen Tabelle durch die entsprechenden Werte der vorlaufenden Tabellen unter Umständen recht genau berechnet werden können, wenn solche zur Schätzung heranzuziehenden Werte nicht gesperrt wurden. Entsprechendes gilt auch in umgekehrter Richtung, wo man offene Werte der aktuellen Tabelle zur Berechnung entsprechender gesperrter Werte der Vorperiodentabelle verwenden kann. – Derartige Schätzverfahren werden ja in der amtlichen Statistik zur Ermittlung von Antwortausfällen schon seit langem eingesetzt. – Andererseits können bei der querschnittsmäßigen Bearbeitung der aktuellen Zeitreihentabelle andere Sperrmuster entstehen als in den Vorperiodentabellen, weil durch natürliche Fluktuationen (Geschäftsaufgaben, Neugründungen) Sperrpositionen zur Sicherung sensibler Tabellenfelder wegfallen oder neu hinzutreten.

Es genügt eben nicht, nur die Sperreintragungen der Vorperiode zu übernehmen (so genanntes Durchstechen). Auch die Übernahme von Vorperiodensperrungen und anschließende Ergänzung durch weitere Sperrungen, die sich bei der querschnittsmäßigen Bearbeitung der aktuellen Tabelle ergeben, führt zu unüberwindbaren Schwierigkeiten:

1. Neueintragungen von Sperrvermerken können weiterhin aus den Vorperiodenwerten durch Schätzung der „Antwortausfälle“ ermittelt werden.

2. Durch die Beibehaltung von Sperrungen aus den Vorperiodentabellen und Ergänzung durch neu hinzuzufügende Sperrvermerke nimmt die Anzahl der zu sperrenden Tabellenwerte von Periode zu Periode ständig zu, niemals ab, so dass schließlich kaum noch zu veröffentlichende Werte in der gerade behandelten Zeitreihentabelle übrig bleiben.

Das Problem der Sicherung von Zeitreihen ist lange bekannt; es gibt bisher keine befriedigende hinreichende Lösung. Die exakte Lösung des Problems besteht natürlich in der Berücksichtigung der Schätzfehler, die aufgrund des vom externen Tabellennutzer verwendeten statistischen Schätzverfahrens bei der Berechnung der geheimen Werte aus den Vorlauftabellenwerten anfallen (vgl. 3.2.3.1). Dieses Vorgehen ist allerdings recht aufwendig, zumal bei unterschiedlichen geheimen Werten ganz unterschiedliche Schätzintervalle auftreten können - wo man schon meist nicht bereit ist, bei der Berücksichtigung von Vorwissen außer einem konstanten relativen Schätzfehler f auch noch davon abweichende Schätzfehler zu berücksichtigen.

Gängige Verfahren sind bisher die separate querschnittsmäßige Behandlung mit einschlägigen Geheimhaltungsverfahren, d.h. ohne Berücksichtigung des zeitlichen Zusammenhanges, d.h. auch ohne Berücksichtigung von Schätzfehlern aus Zeitreihen-Schätzungen, und selektives Durchstechen in Verbindung mit querschnittsmäßiger Restsicherung, wobei eine mehr oder weniger unvollständige Übertragung der Vorperiodensperrvermerke erfolgt.

5.3.2.1 Gewichtung nach Sperrpositionen der Vorperiodentabelle

Das zuletzt genannte Vorgehen, eine Kombination aus teilweisem Durchstechen und querschnittsmäßiger sekundärer Geheimhaltung, lässt sich nun mit Hilfe der externen Gewichtung auf einfache Weise formalisieren und dadurch in ein allgemein anwendbares EDV-Verfahren überführen. Dazu bietet sich an, die im Datenmaterial der Vorperiodentabelle gegebene Verteilung der Sperrvermerke (primäre wie sekundäre) auf den aktuellen Datenbestand in Form geeigneter Gewichte abzubilden. Dies geschieht zweckmäßig nach dem bewährten Vorbild der Randsummengewichtung (siehe Punkt 5.2.1), indem hier die in der Vorperiodentabelle offenen Werte in der aktuellen Tabelle besonders hoch gewichtet werden, damit diese Werte bei der anschließenden querschnittsmäßigen Bearbeitung mit dem Geheimhaltungsverfahren nach Möglichkeit offen bleiben.

Weist der Sperrschlüssel eines Tabellenfeldes der Vorperiode also einen offenen Wert aus, so wird das Gewicht im Gewichtsfeld des aktuellen Tabellenbestandes mit einem hohen positiven reellen Zahlenwert, etwa größter Klassenwert * Anzahl der Quadereckwerte pro Sicherungsquader, versehen. – Der Gewichtswert sollte immer wesentlich größer als die Summe ungewichteter offener Werte eines Quaders sein, damit das Produkt aus kleinstem Klassenwert und großem Gewichtswert nicht kleiner als die Summe ungewichteter offener Werte eines Quaders ist, was sonst zur unerwünschten Bevorzugung eines höher gewichteten offenen zu sperrenden Tabellenwertes gegenüber einem ungewichteten Wert als Sicherungspartner führen könnte. - Weil bei negativer Gewichtung zur Erzwingung von Sperrungen in vorgegebenen Tabellenbereichen durch die Umkehrung des Vorzeichens eine Umkehrung der Ordnung entstehen würde, ist die oben beschriebene positive Gewichtung bei in der Vorperiode offenen Werten einer negativen Gewichtung bei in der Vorperiode gesperrten Tabellenwerten unbedingt vorzuziehen.

Ungewichtete Tabellenwerte werden bei der querschnittsmäßigen Bearbeitung mit dem Geheimhaltungsprogramm GHQUAR, GHMITER oder QUIT nun besonders bevorzugt gesperrt, wodurch das Durchstechen realisiert wird. Dennoch werden nur so viele Tabellenwerte im aktuellen Bestand gesperrt, wie es die Sicherung mit Intervallschutz erfordert; es erfolgt kein bedingungsloses Durchstechen.

5.3.2.2 Gewichtung nach relativen Schätzfehlern

Eine Verfeinerung der Sperrpositionen-Auswahl lässt sich noch durch die Gewichtung nach der Schätzgenauigkeit, mit der ein Tabellenwert aus offenen Werten anderer Zeitreihentabellen berechnet werden kann, erreichen. Als Gewichtswert kommt dabei das Quadrat des relativen Schätzfehlers bzw. der Reziprokwert davon in Betracht, je nachdem ob verhindert werden soll, dass geheim gehaltene Vorperiodenwerte aus aktuellen Werten zu genau geschätzt werden können, oder ob der aktuelle Tabellenwert, wenn er denn gesperrt würde, aus den Vorperiodenwerten zu genau berechenbar wäre.

Als Maß für den relativen Schätzfehler bietet sich der absolute Betrag der Abweichung des aktuellen vom Vorperiodenwert bezogen auf den aktuellen Wert an; ist der aktuelle Wert Null, so ist statt dessen nur das Quadrat des Abweichungsbetrags (Betrag des Vormonatswerts) bzw. dessen Kehrwert als Gewicht in der Quadersumme zu verwenden. Dieser Fehlerabschätzung liegt die Erfahrung zu Grunde, dass sich Antwortausfälle bei kurzzeitig aufeinanderfolgenden Zeitreihentabellen sehr gut durch Übertragung von Vorperiodenwerten schätzen lassen. Selbstverständlich können aber auch andere Fehlermaße zur Gewichtung herangezogen werden, wie z.B. der relative

Standardfehler der Verhältnisschätzung, wenn bei dem zu sichernden Zeitreihentyp zur Schätzung von Antwortausfällen eine Verhältnisschätzung erfahrungsgemäß genauere Ergebnisse liefert.

Bei der Vergabe der Gewichte in der aktuellen Zeitreihentabelle hat man prinzipiell zwei Fälle zu unterscheiden:

1. Schätzfehlergewichtung zum Schutze geheimer Vorperiodenwerte

   Die Gewichtung eines aktuellen Tabellenwertes muss so erfolgen, dass dieser Wert bevorzugt gesperrt wird, wenn mit seiner Hilfe geheime Vorperiodenwerte besonders genau berechnet werden können, wenn also der relative Schätzfehler des hier zu schützenden Vorperiodenwertes besonders klein ist. Als in die aktuelle Tabelle einzutragendes Gewicht ist daher das relative Abweichungsquadrat des aktuellen vom Vorperiodenwert zu wählen, weil das danach anzuwendende sekundäre Geheimhaltungsverfahren den betreffenden Tabellenwert um so eher sperrt, je kleiner sein Gewicht ist. Diese Gewichtung betrifft also nur Werte der aktuellen Tabelle, deren Vorperiodenwerte gesperrt sind.

2. Reziproke Schätzfehlergewichtung zum Schutze geheimer Werte in der aktuellen Tabelle

   Wenn auf Grund des Sperrmusters zu schützender Werte in der aktuellen Tabelle Sperreintragungen vorzunehmen sind, denen in der Vorperiode noch offene Werte gegenüberstehen, so besteht die Gefahr, dass die geheim gehaltenen aktuellen Werte aus der Vorperiodentabelle berechnet werden können. Dagegen schützt in gewissen Grenzen eine Gewichtung des aktuellen Wertes mit dem Reziprokwert des relativen Schätzfehlerquadrates: Aus den Vorperiodenwerten besonders genau zu berechnende aktuelle Tabellenwerte werden dann auf Grund ihres großen Gewichts (als Kehrwert eines kleinen Schätzfehlerquadrates) bei der Auswahl von Sekundärsperrkandidaten weitgehend gemieden. Von dieser Art der Gewichtung betroffen sind also nur solche Werte der aktuellen Tabelle, deren Vorperiodenwerte offen sind.

Bei der Sicherung von Zeitreihentabellen unter Berücksichtigung der zeitlichen Abhängigkeit ihrer Tabellenwerte von den Vorperiodenwerten sollte bei der Bearbeitung mit dem Sekundär-Geheimhaltungsverfahren unbedingt von der externen Gewichtung nach Sperrpositionen der Vorperiodentabelle Gebrauch gemacht werden (Punkt 5.3.2.1). Eine weitergehende Differenzierung, insbesondere bei hinsichtlich der Schätzfehler zur Schätzung von „Antwortausfällen“ sehr heterogenem Datenmaterial, kann dann noch eine Gewichtung mit Schätzfehlern der Gewichtung nach Punkt 5.3.2.1 überlagert werden, indem die bereits eingetragenen Gewichtswerte (auch Gewichte = 1) mit dem Schätzfehlergewicht (Unterpunkte 1 und 2) multipliziert werden.

Durch die anforderungsgerechte Übertragung von Vorperiodensperrungen in Zeitreihentabellen mit Hilfe der externen Gewichtung kann dem Statistiker ein nunmehr objektives Verfahren an die Hand gegeben werden, mit dem er die starke Abhängigkeit in kurzen zeitlichen Perioden aufeinander folgender Statistikerhebungen (monatliche, vierteljährliche) bei der Durchführung der (sekundären) Geheimhaltung in befriedigender Weise berücksichtigen kann.

5.3.3 Instantane Gewichtung

Viele Nutzer wünschen sich eine Auswahl von Sperrkandidaten, die sich am Abstand vom jeweiligen zu schützenden Pivot-Element in der als metrischer Raum betrachteten Tabelle orientiert. Als Beispiel führen sie nach Größenklassen gegliederte Daten an, bei denen ein geheimer Wert einer bestimmten Größenklasse nach Möglichkeit durch einen anderen geheimen Wert in einer benachbarten Größenklasse geschützt werden sollte. – Der hier auf dem geometrischen Abstand beruhende Nachbarschaftsbegriff ist von dem durch Paare von Werten, die zur selben Quadersumme beitragen, wohl zu unterscheiden. –

Ein anderes Beispiel ist die Umbuchung von Fällen innerhalb eines Quaders (4.1 Quaderverfahren zur Werteverfälschung), bei der Sicherungsquader zu bevorzugen sein werden, deren geometrischer Abstand aller im Quader benachbarter Werte besonders klein ist, weil die dann umgebuchten Berichtenden sich hinsichtlich ihrer erhobenen Merkmale besonders wenig voneinander unterscheiden, sie also besser in das neue Tabellenfeld hineinpassen. Beim ersten Beispiel wird die Differenzierung durch die Abstandsfunktion sich nur auf die Größenklassengliederung beziehen müssen, beim zweiten Beispiel auf alle Gliederungen einer Tabelle.

Die Bevorzugung von geometrisch benachbarten Werten beim Sperrprozess lässt sich noch am einfachsten mit Hilfe des Summenkriteriums durch eine instantane Gewichtung der Werte des auszuwählenden Quaders realisieren, bei der die zu sperrenden Werte in der Quaderwertesumme mit einem vom Nutzer vorgebbaren, die Tabellengeometrie betreffenden Abstandsmaß gewichtet werden. Nach dem Summenkriterium werden dann solche Quader als Sicherungsquader zum Schutze geheimer Tabellenwerte bevorzugt, deren Quaderwerte besonders kleine Gewichtsfaktoren haben, die also im Sinne dieses Abstandsmaßes alle zueinander und insbesondere zu dem zu schützenden Pivot-Tabellenfeld besonders nahe benachbart sind.

Um die instantane Gewichtung in einem EDV-Programm wie GHQUAR zu realisieren, ist für jeden Wert Q des zum Schutze des Pivots G auszuwählenden Quaders die Gewichtsfunktion W(Q,G) anzusetzen:

$$\mathbf{W(Q,G) = |q_1-g_1|^{p}{}_1 + |q_2-g_2|^{p}{}_2 + ... + |q_n-g_n|^{p}{}_n}$$

Dabei bezeichnen

n	=	Tabellendimension
$q_1, q_2, ..., q_n$	=	Koordinaten des Quaderwertes Q
$g_1, g_2, ..., g_n$	=	Koordinaten des Pivot-Wertes G
$p_1, p_2, ..., p_n$	=	Potenzen der absoluten Beträge der Koordinatendifferenzen $\|q_i-g_i\|$, i=1,2,...,n, mit denen diese in das Abstandsmaß eingehen.

Die große Mannigfaltigkeit dieser Gewichtung lässt sich mit Hilfe nachstehender Hinweise etwas überschaubarer machen; in Zweifelsfällen hilft eine Vorabauswertung:

Sollen alle Gliederungskriterien hinsichtlich ihres Abstandsmaßes gleichstark in die Gewichtung eingehen, so ist

$p_1 = p_2 = ... = p_n = p$

zu setzen; insbesondere ergibt p = 2 das Quadrat des Euklidischen Abstandsmaßes.

Positive Werte der p_j erhöhen das Gewicht (verringern also die Sperrwahrscheinlichkeit von Werten mit größeren Koordinatendifferenzbeträgen), negative p_j verkleinern es. Potenzen mit Wert $p_j = 0$ machen die Gewichtung von den betreffenden Gliederungsmerkmalen unabhängig. Ist beispielsweise nur p_1 von Null verschieden, so ist die Gewichtsfunktion $W = |q_1-g_1|^{p}_1 + n - 1$.

Für jeden Wert Q des zum Schutze des Tabellenwertes G auszuwählenden Quaders wird der zugehörige Gewichtswert W(Q,G) berechnet und sein Klassenwert dann mit W(Q,G) gewichtet (multipliziert) in die Quaderwertesumme des Summenkriteriums eingetragen (andere Gewichtsfaktoren bleiben davon unberührt). Alle für die Berechnung der instantanen Gewichtsfunktionswerte benötigten Koordinatenwerte (Ausprägungen der Gliederungsmerkmale) q_i, g_i, i = 1, 2, ..., n, liegen in einem EDV-Programm zum Quaderverfahren am Ort der Ausführung zugriffsbereit vor. Die vom Nutzer vorzugebenden Potenzen p_i, i = 1, 2, ..., n, müssen noch per „Steuerkarte“ eingelesen werden.

6. Sicherung von Tabellen mit gemeinsamen Aggregaten

6.1 Tabellenübergreifende Geheimhaltung

In der "statistischen Praxis" hat man es häufig mit mehreren - auch mehrfach durch Zwischensummen unterteilten - Tabellen zu tun, die einander überlappen, d. h. die gewisse Aggregate gemeinsam haben: So kann beispielsweise der regional gegliederte steuerbare Umsatz einmal nach Rechtsformen der Betriebe, ein anderes Mal nach Beschäftigtengrößenklassen oder auch nach wirtschaftlicher Systematik "heruntergebrochen" werden. Gemeinsam haben diese Tabellen die nur regional gegliederte Summentabelle.

Bei der Tabellensicherung träte hier kein neues Problem auf, wenn es gelänge, die Überlappungsbereiche beim Sperren von Werten zu meiden. Untersuchungen im Zusammenhang mit der Geheimhaltung der Handwerkszählung 1995 haben aber gezeigt, dass Sperrungen in Überlappungsbereiche prinzipiell nicht auszuschließen sind. Daraus ergibt sich als zwingende Notwendigkeit, dafür zu sorgen, dass mehreren Einzeltabellen gemeinsam angehörende Aggregate in allen diesen Einzeltabellen den gleichen Geheimhaltungsstatus haben. Für obiges Beispiel bedeutet das, dass die nur noch regional gegliederten Gesamtsummenwerte in jeder der drei Einzeltabellen, der Rechtsformen-, der Beschäftigtengrößenklassen- und der nach wirtschaftlicher Systematik gegliederten Tabelle, zumindest den gleichen Geheimhaltungsvermerk tragen. Außerdem müssen auch Schätzintervalle übertragen werden, die bei nachfolgenden Betrachtungen jedoch noch unberücksichtigt bleiben.

Für die Sicherung solcher voneinander abhängiger Tabellen kommt daher nur eine gemeinsame Bearbeitung durch gegenseitigen Abgleich in Frage, ganz analog zum Abgleich der Untertabellen. Das bedeutet, dass alle zu einem Pool aneinander abzugleichender Tabellen gehörenden Einzeltabellen i.d.R. gleichzeitig veröffentlicht werden. Eine später erstellte Veröffentlichungstabelle, die Überlappungen mit vorhergehenden, bereits veröffentlichten Tabellen hat, kann nur dann gesichert werden, wenn der Abgleich mit allen in Frage kommenden "Vorgängertabellen“ ausschließlich Sperrungen in der „Nachzüglertabelle“ hervorbringt, sonst nicht.

6.1.1 Verschränkte Einzeltabellenverarbeitung

Auch für diesen Abgleich von einzelnen Veröffentlichungstabellen ist im LDS NRW ein EDV-Programm entwickelt worden, das Programm GHMITER zur iterativen Durchführung der Geheimhaltung bei überlappenden Statistiktabellen. Eine Anwendung des Programms auf Realdaten zeigt das zweite Beispiel des Anhangs A. Das hier zunächst diskutierte Verfahren der Einzeltabellenverschränkung wird durch den Untertabellenabgleich nahegelegt, ist vergleichsweise einfach darzustellen, eignet sich aber nur für nicht zu große Einzeltabellen (< 100.000 Tabellenfelder).

Den Eingabe- und Arbeitsbestand für dieses iterative Abgleichsverfahren erhält man, indem man eine n-dimensionale Tabelle nach der anderen in den Datenbestand überträgt und mehrfach vorkommende Sätze löscht (Verschränkung der Einzeltabellen). Dabei werden alle Gliederungsmerkmale der Einzeltabellen als neue Pooltabellenmerkmale in den Arbeitsbestand übernommen. Der Überlappungsbereich zeigt sich dadurch, dass dort mehr Gliederungsmerkmale auftreten als in jeder abzuspeichernden Einzeltabelle, die zu dieser Überlappung beitragen; sonst findet man im Gesamtbestand immer die Ausprägungen von Gliederungsmerkmalen nur einer Einzeltabelle. Die Struktur der Gliederungskriterien wird, wie bei der Bearbeitung der Einzeltabellen auch, in einer weiteren Datei der Schlösser nachgehalten.

Der iterative Abgleich erfolgt dann so, dass eine Statistiktabelle nach der anderen als Projektion aus dem Datenbestand in das Geheimhaltungsprogramm übernommen wird. Nach der Bearbeitung jeder Einzeltabelle werden die veränderten Wertartschlüssel temporär gespeichert und in den nachfolgenden Iterationsschritten mitberücksichtigt. Sind alle Tabellen abgearbeitet, beginnt das Verfahren von neuem, bis nach einem vollen Durchlauf keine neuen Sperrvermerke (Wertart) in den Überlappungsbereich zurückgeschrieben werden müssen. Der gemeinsame Abgleich aller Tabellen erfolgt hier einfach durch das Überschreiben der Wertart im Überlappungsbereich.

Bei der Bearbeitung einander überlappender Tabellen können Übersperrungen auftreten, weil bei jedem neuen Durchlauf auch die Sekundärsperrungen geprüft und ggf. gesichert werden. Das führt bei Intervallschutz und insbesondere bei Einzelangaben oft zu weiteren Sperrungen, weil zwar jeder primär geheime Wert durch die nur zu seinem Schutz gesetzten Sekundärsperrungen vollkommen gesichert ist, nicht unbedingt aber umgekehrt die Sekundärsperrungen durch Einzelangaben oder durch andere primäre Sperrungen. Um solche durch iterativen Tabellenabgleich bedingten Übersperrungen zu vermeiden, werden ab dem zweiten Iterationsschritt die während des jeweils vorangegangenen Durchlaufs neu gesetzten sekundär geheimen Werte in einer besonderen Klassenstaffel (vgl. 5.2.1) gesammelt; der laufende Iterationsschritt sichert dann nur noch diese Werte, während die anderen, bereits vollständig geschützten Tabellenwerte unbehelligt bleiben.

Dieses Verfahren ist auch bei Geheimhaltung mit Intervallschutz zweckmäßig, wenn kein iterativer Abgleich der Schutzintervalle vorgesehen ist (siehe zweite Anmerkung zu 3.1.2). Bei Übertragung von Schätzintervallen mit Intervallabgleich werden in der Regel bereits bestimmte Schutzintervalle weiter eingeengt (vgl. 3.2.3) und zwar unabhängig vom Zeitpunkt des Eintrags der Sperrung. Dadurch verändert sich aber der Geheimhaltungsstatus der betroffenen Werte, so dass bei jedem Iterationsschritt immer alle geheimen Werte hinsichtlich ihres Intervallschutzes überprüft werden müssen; Übersperrungen sind unter solchen Umständen unvermeidbar.

Die oben dargestellte Tabellenorganisation zur gemeinsamen Bearbeitung mit dem Geheimhaltungsverfahren ist sehr allgemein anwendbar. Sie beschränkt sich nicht nur auf Tabellen mit gemeinsamen Randsummentabellen, sondern ist auch einsetzbar, wenn „innere“ Tabellenteile, also auch niedrigere Aggregate wie Zwischensummen zum Überlappungsbereich gehören. So kann man beispielsweise zu Gunsten der Veröffentlichung einer sehr feinen Gliederung in Bezug auf das erste Gliederungskriterium eine wesentlich höhere Verdichtung bezüglich des zweiten Gliederungsmerkmals einer Tabelle in Kauf nehmen wollen. Um dennoch auf eine detaillierte Darstellung der Daten auch bezüglich der zweiten Merkmalsgliederung nicht verzichten zu müssen, kann man auch die in umgekehrter Weise verfeinerte bzw. vergröberte Tabelle veröffentlichen, bei der nun das erste Gliederungskriterium das vergrö-

berte, das zweite das verfeinerte ist. Diese beiden Tabellen haben gemeinsame Aggregate in ihrem Zwischensummenbereich und müssen, wenn beide veröffentlicht werden sollen, auch gemeinsam mit einem Geheimhaltungsverfahren für überlappende Tabellen bearbeitet werden, das einen Tabellenabgleich vornimmt. Auch das vermag das o.g. Verfahren zu bewerkstelligen (siehe Benutzeranleitung GHMITER).

Die Beschäftigung mit überlappenden Tabellen kann noch aus einem anderen Grunde zweckmäßig sein: Betrachtet man zunächst die Ausgangsdaten einer Statistik, die in der Regel durch die Erhebungsbögen gegeben sind, so können diese Daten nach einer großen Anzahl von Merkmalen gegliedert und auch aggregiert werden. Das Ergebnis ist eine sehr umfangreiche, meist viele Millionen Datenfelder umfassende, hochdimensionale Statistiktabelle, die als Ganzes niemals veröffentlicht wird. Es liegt daher nahe, nicht die alles umfassende Tabelle abzuspeichern und hinsichtlich der Geheimhaltung zu bearbeiten, sondern nur diejenigen Tabellenteile, die von öffentlichem Interesse sind. Dieses Vorgehen verspricht eine erhebliche Speicherplatz- und Rechenzeitersparnis (CPU-Zeit). Die danach verbleibenden Daten sind als Projektionen aus dem Gesamtdatenbestand meist hochverdichtete niedrigdimensionale, einander überlappende Tabellen, die die höheren Aggregate häufig gemeinsam haben. Auch auf solche Tabellen ist obiges Abgleichsverfahren anwendbar.

Nach dem soeben aufgezeigten Muster kann man auch große, mehrfach durch Zwischensummen unterteilte Statistiktabellen in einzelne Teiltabellen zerlegen, von denen jede selbst wieder mehrfach durch Zwischensummen untergliedert sein kann, und diese Tabellenteile als einen Pool überlappender Tabellen auffassen, der dann mit obigem Iterationsverfahren gesichert werden muss. Diese Art der Tabellenzerlegung ist vor allem dann angezeigt, wenn gewisse größere Teiltabellen in keiner Veröffentlichung auftreten. Man hat damit eine Alternative zur Aussparung von Tabellenteilen mittels externer Gewichtung, die u.U. einfacher zu handhaben ist als die Einführung externer Gewichte.

Prinzipiell könnte man mit der Zerteilung einer Statistiktabelle fortfahren, bis als Teiltabellen des Pools überlappender Tabellen nur noch die Untertabellen der ursprünglichen Statistik zu finden sind. Das Ergebnis der Geheimhaltung mit obigem Iterationsverfahren wäre das gleiche wie das gemäß Abschnitt 1 mit hierarchischem Untertabellenabgleich erhaltene, wenn man in obiger Iteration alle Untertabellen nach absteigenden Aggregationsniveaus abarbeitete. Um Übersperrungen dabei zu vermeiden, muss außerdem gemäß obiger Bemerkungen dafür gesorgt werden, dass bei der Iteration Sekundärsperrungen nur dann gesichert werden, wenn sie durch Untertabellenabgleich eingetragen wurden .

Es ist aber nicht ratsam, den Untertabellenabgleich extern zu steuern, weil dabei viel Kanalrechenzeit und damit eine beträchtliche Gesamtrechenzeit hinzunehmen wäre. Diese Anmerkung verdeutlicht aber den direkten Zusammenhang, der zwischen dem „internen“ Untertabellenabgleich und dem „externen“ Abgleich einzelner einander überlappender Statistiktabellen besteht. Im Übrigen ist beiden Verfahren gemeinsam, dass sie für die Sicherung geheimer Tabellenwerte nur notwendig, nicht jedoch hinreichend sind, wie im Abschnitt 6.2.1 gezeigt wird.

6.1.2 Einzeltabellen in einem übergeordneten Tabellenraum

Als überlappende Tabellen werden hier Veröffentlichungstabellen einer einzelnen Statistik betrachtet. Bei umfangreichen Statistiken können auch umfangreiche Einzeltabellen zu bearbeiten sein, die bei jedem Iterationsschritt aus ihrer Verschränkung gemäß 6.1.1 zu extrahieren wären. Dieses Verfahren wäre zu rechenzeitaufwendig. Im Folgenden werden daher die Einzeltabellen als Hauptspeichertabellen getrennt geführt; die Festlegung ihres Überlappungsbereiches erfordert eine eingehendere Betrachtung des Aufbaus der zu sichernden aggregierten Daten. Dieses Verfahren ist mit dem EDV-Programm GHMITER ab Version 2 und mit QUIT realisiert.

Eine Statistik ist die Gesamtheit der (z.B. durch Meldebögen) erhobenen Daten, die nach einer Teilgesamtheit dieser Daten, der Gesamtheit aller r Gliederungsmerkmale (r steht für rigoros), gegliedert ist, d.h. die in einem von diesen r Gliederungsmerkmalen aufgespannten Raum, dem kartesischen Produkt dieser Gliederungsmerkmale, eingeordnet werden können. Die nicht als Gliederungsmerkmale angesehenen Merkmale sind die Werte der Statistik. Zu einem r-Tupel von Gliederungsmerkmalen können $w \geq 1$ Werte gehören, hier wird nur ein einziger Wert – z.B. als führendes Merkmal - bei der Geheimhaltung behandelt, die anderen Merkmale können dann entweder einzeln nacheinander dem Geheimhaltungsprozess unterworfen oder mit demselben Sperrvermerk wie das führende Merkmal belegt werden (Durchstechen).

Eine Einzel- oder Veröffentlichungstabelle bezeichnet eine nach wenigen Gliederungsmerkmalen gegliederte Projektion der Statistik auf den von den Gliederungsmerkmalen der Einzeltabelle aufgespannten Raum, in dem spezielle, nicht zur Einzeltabellengliederung gehörige Gliederungsmerkmale festgehalten werden oder/und in dem über gewisse andere Gliederungen der Statistik summiert wird. Außerdem können diese Einzeltabellengliederungen aus denen der Statistik durch Weglassen von Kategorien verkürzt oder durch Zufügung von Summen erweitert worden sein. Häufig sind Einzeltabellen durch Zwischensummen untergliedert; zu jedem Gliederungsmerkmal haben diese Einzeltabellen aber immer nur eine (Rand-)Summe höchster Aggregation.

Veröffentlicht werden immer nur niedrigdimensionale, d.h. hochaggregierte Teilgesamtheiten einer Statistik, die aus den sogenannten Summensätzen aufgebaut werden, die Menge sämtlicher Summensätze, die für alle Veröffentlichungstabellen ausreicht, erhält die Bezeichnung Veröffentlichungsdaten. Sie können in demjenigen p-dimensionalen Veröffentlichungsraum dargestellt werden, der von den p Gliederungsmerkmalen aller Veröffentlichungstabellen der Statistik aufgespannt wird (p steht für publik), $p \leq r$. Im Veröffentlichungsraum sind n_i-dimensionale Einzeltabellen durch ihre n_i Gliederungsmerkmale, $n_i \leq p$, sowie durch $p\text{-}n_i$ vorgegebene, die Einzeltabelle im p-dimensionalen Raum fixierende Gliederungsausprägungen, ihre Pseudoindizes, charakterisiert, i = 1, 2, ..., I, I = Anzahl der Veröffentlichungstabellen.

Der Überlappungsbereich von Veröffentlichungstabellen einer Statistik ist ganz allgemein die Gesamtheit der Aggregate, die mehrere dieser Tabellen gemeinsam haben. Dabei können unterschiedlich viele Veröffentlichungstabellen einander überlappen. Als ein Bereich s-facher Überlappung, $s \geq 2$, wird daher eine Gesamtheit von Aggregaten bezeichnet, die jeweils s Einzeltabellen gemeinsam angehören; er ist gekennzeichnet durch eine Gesamtheit von p-

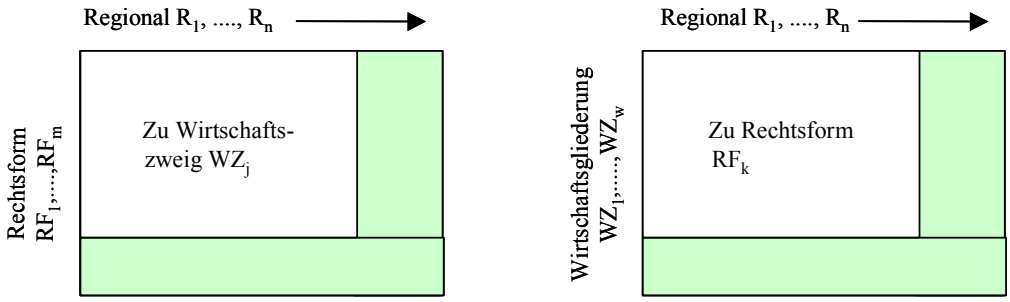
Tupeln, den Ausprägungen der Gliederungsmerkmale, die die Aggregate des Bereichs s-facher Überlappung im Veröffentlichungsraum gemeinsam haben und die ihn dadurch festlegen.

Für die Weitergabe von Daten und damit insbesondere für die primäre wie auch die sekundäre Geheimhaltung einzig relevanter Bezugsraum ist der Veröffentlichungs- oder auch Projektionsraum. Alle die tabellarische Anordnung von Veröffentlichungsdaten innerhalb von Einzeltabellen wie auch in Bezug auf mehrere Einzeltabellen zueinander betreffende Eigenschaften lassen sich im Projektionsraum geometrisch konkretisieren. Dies gilt speziell auch für die Überlappungsbereiche, die hier – wie oben bemerkt – durch diejenigen p Gliederungsausprägungen gekennzeichnet sind, deren Aggregate zwei oder mehr als zwei Einzeltabellen angehören ($s \geq 2$).

Der eigentliche Abgleichsprozess, der dafür sorgt, dass mehreren Einzeltabellen gemeinsame Werte alle den gleichen Geheimhaltungsstatus haben, erfolgt immer nur durch paarweisen Abgleich von Einzeltabellen. Die Abgleiche von Mehrfachüberlappungen werden demnach beim Abgleichsprozess in lauter paarweise Abgleiche zerlegt. Zur Veranschaulichung genügt also die Betrachtung der Schnittmenge zweier überlappender zweidimensionaler Tabellen in einem dreidimensionalen Veröffentlichungs- bzw. Projektionsraum.

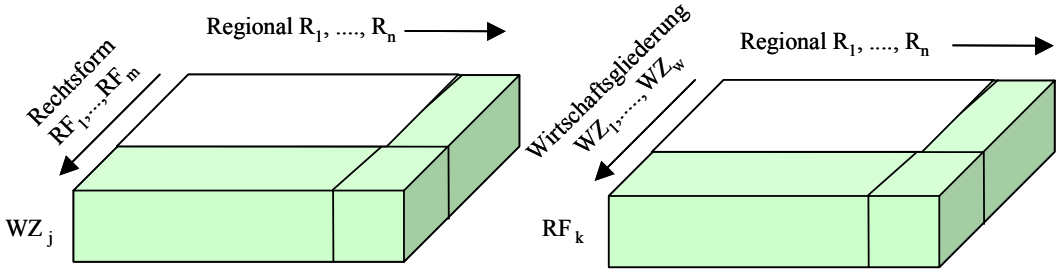
Schema der Einfügung von Einzeltabellen in einen Veröffentlichungsraum

a) Gegebene Einzeltabelle (Zwischensummen nicht eingezeichnet) :

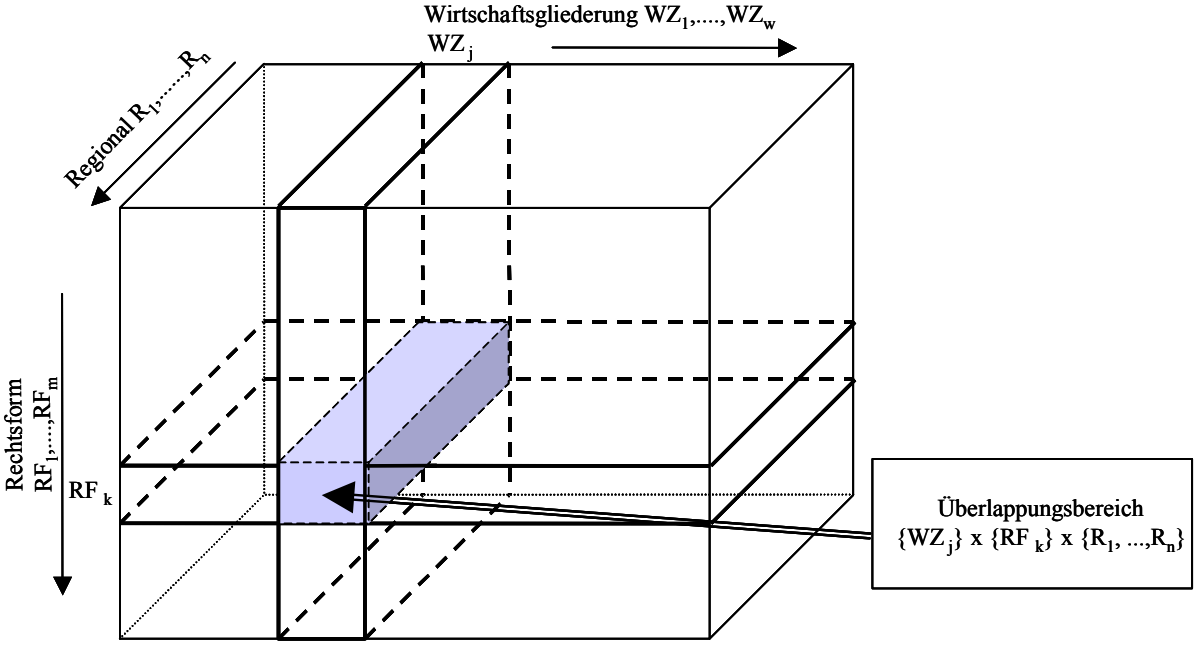


b) Einordnung in die Gesamttabelle :

ba) Auffassung der Tabellenparameter als zusätzliche Indizes :



bb) Einordnung der indexerweiterten Einzeltabellen in eine (fiktive) Gesamttabelle :



Die Veranschaulichung des Überlappungsproblems macht deutlich, wie im Allgemeinen bei der Behandlung von mehreren aus einer Statistik gewonnenen Einzeltabellen mit dem Verfahren der sekundären Geheimhaltung vorzugehen ist:

- Kennzeichnung aller Einzeltabellen durch sämtliche p Indizes des Projektionsraums; dabei sind auch diejenigen Indizes zu berücksichtigen, die bei der Projektion jeder Einzeltabelle aus dem Veröffentlichungsraum, als Pseudoindizes (als Parameter der Einzeltabelle) auftreten.

- Bestimmung des Überlappungsraumes jedes Paares überlappender Tabellen als Gesamtheit der Daten, deren Index-p-Tupel (aus Pseudo- wie auch aus laufenden Indizes der beiden Tabellen) in beiden Einzeltabellen jedes Paares vorkommen.

- Abspeicherung dieser Indizes als Überlappungsindizes zwecks Kennzeichnung der in den Überlappungsbereich einzutragenden Sperrvermerke.

Anmerkung

Die Indizes WZ_j, RF_k können selbst bereits aus mehreren Indizes bestehen , z. B

$$WZ_j = \{ WZ_{j1} , ... , WZ_{j\,i} \}$$

und $RF_k = \{ RF_{k1} , ... , RF_{k\,v} \}$,

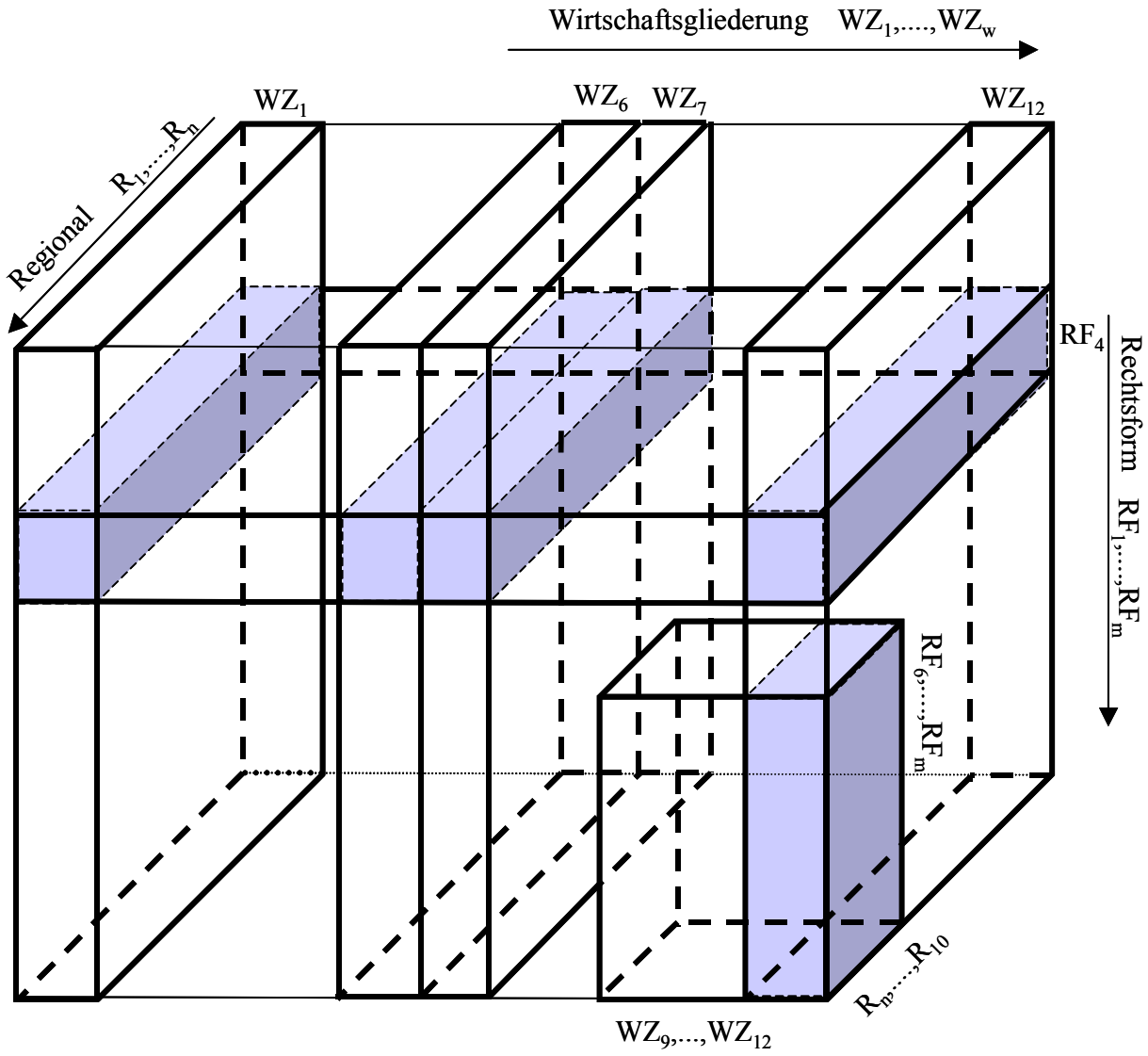
dann ist der Überlappungsbereich in diesem Fall

$$\{ WZ_{j\,1} , ... , WZ_{j\,i} \} \times \{ RF_{k\,1} , ... , RF_{k\,v} \} \times \{ R_1 , ... ,R_n \}$$

Der Überlappungsbereich des nachfolgenden Schaubildes ist beispielsweise

$$\{ WZ_1 , WZ_6 , WZ_7 , WZ_{12} \} \times \{ RF_4 \} \times \{ R_1 , ... ,R_n \} \cup$$

$$\{ WZ_{12} \} \times \{ RF_6 , ... , RF_m \} \times \{ R_{10} , ... ,R_n \}$$



6.2 Rückführung „überlappender“ auf „vollständige“ Tabellen

6.2.1 Rückrechenbarkeit aneinander abgeglichener Untertabellen

Wie bemerkt, wurde bisher auch der gegenseitige Untertabellenabgleich in Bezug auf die Summensperrungen dadurch erreicht, dass die gesamte Untertabellenhierarchie in mehreren Iterationsschritten so lange durchlaufen wurde, bis keine weiteren Sekundärsperrungen mehr auftraten. Dieses Vorgehen ist zwar notwendig, nicht jedoch hinreichend für die Sicherung der Gesamttabelle. Dazu betrachte man folgendes Gegenbeispiel:

Abb. 6.1

Rückrechenbarkeit über mehrere in sich sichere und aneinander abgeglichene Untertabellen

X_1	0	X_1	X_2	0	X_2	X_3	0	X_3	10
X_4	0	X_4	0	X_5	X_5	0	X_6	X_6	20
20	0	20	X_2	X_5	X_2+X_5	X_3	X_6	X_3+X_6	30
0	0	0	X_7	0	X_7	X_8	0	X_8	40
0	0	0	0	X_9	X_9	0	X_{10}	X_{10}	20
0	0	0	X_7	X_9	X_7+X_9	X_8	X_{10}	X_8+X_{10}	60
20	0	20	20	10	30	15	25	40	90

Es gilt:	$X_1 + X_2 + X_3$	=	10
	$0 + X_7 + X_8$	=	40
	_____		
(1)	$X_1 + (X_2+X_3+X_7+X_8)$	=	50
	$X_2 + X_7$	=	20
	$X_3 + X_8$	=	15
	_____		
(2)	$(X_2+X_3+X_7+X_8)$	=	35
(1) - (2) ergibt:	X_1	=	15

Anmerkung: In dieser Tabelle müssen auch negative Zahlen vorkommen, denn die Randsumme der ersten Zeile (= 10) ist kleiner als der erste Summand ($X_1 = 15$)!

Die Ursache für die über mehrere Untertabellen laufende Rückrechenbarkeit geheimer Werte liegt in der Aufteilung des durch die Summationsvorschriften der Gesamttabelle gegebenen linearen Gleichungssystems zur Berechnung dieser geheimen Werte auf die einzelnen Untertabellen. Bei Summensperrungen sind diese Teilsysteme, deren Untertabellen gemeinsam zu denselben Randsummen beitragen, nicht unabhängig voneinander, müssen also bei der Sicherung auch gemeinsam bearbeitet werden.

Es muss an dieser Stelle ausdrücklich betont werden, dass ganz allgemein die etwaige Rückrechenbarkeit einander überlappender Tabellen, d.h. Tabellen, die gemeinsame Tabellenfelder besitzen, keine Besonderheit des Quaderverfahrens ist, sondern unabhängig vom Untertabellensperr-Algorithmus auftritt. Das ist auch der Grund dafür, dass mit den Methoden der linearen Optimierung gesicherte Untertabellen in der Gesamttabelle keinen hinreichenden Schutz bieten.

6.2.2 Aufstockung der Tabellendimension

Die Zusammenfassung durch gemeinsame Sperrungen gekoppelter Untertabellen bedeutet die Einführung weiterer Gliederungsstrukturen bzw. die Aufstockung der Tabellendimension. Dazu betrachte man folgende mehrfach durch Zwischensummen untergliederte eindimensionale Statistiktabelle.

Abb. 6.2

a_1,a_2,a_3, ...	$\sum_1$	b_1,b_2,b_3, ...	$\sum_2$	- - -	v_1,v_2,v_3, ...	$\sum_m$	$\sum\sum$

Darin werden die Elemente der ersten Aggregationsstufe zu ihren Zwischensummen $\sum_i$ zusammengefasst, die dann - ebenfalls aufaddiert - die Gesamtsumme $\sum\sum$ ergeben.

Aufgrund des Assoziativgesetzes und der Kommutativität der Addition hätte man diese Zusammenfassung zur Gesamtsumme aber auch so vornehmen können, dass zunächst die jeweils ersten Elemente der ersten Aggregationsstufe, also die Elemente a_1, b_1, ...,v_1 zu einer Zwischensumme $\sum^*_1$ aufaddiert würden, dann die zweiten Elemente der ersten Aggregationsstufe, a_2, b_2, ...,v_2 zu $\sum^*_2$ usw., um dann die Zwischensummen $\sum^*_j$ - die zweite Aggregationsstufe nach dieser neuen Gliederung - zur Gesamtsumme $\sum\sum$, der dritten Aggregationsstufe, zusammenzufassen.

Dazu schreibt man zweckmäßig die nach der zuletzt genannten Gliederung zu addierenden Werte untereinander und erhält so die in der Abb. 6.3 gezeigte zweidimensionale Tabelle, die nicht mehr durch Zwischensummen untergliedert ist und die im Folgenden als vollständig bezeichnet werden soll.

Abb. 6.3

$a_1, \quad a_2, \quad a_3, \quad - \quad - \quad -$	$\sum_1$
$b_1, \quad b_2, \quad b_3, \quad - \quad - \quad -$	$\sum_2$
| | |	|
| | |	|
| | |	|
$v_1, \quad v_2, \quad v_3, \quad - \quad - \quad -$	$\sum_m$
$\sum^*_1, \sum^*_2, \sum^*_3, \quad - \quad - \quad -$	$\sum\sum$

Bei der Umstellung kann es vorkommen, dass die Anzahl der Kategorien bezüglich der aufzustockenden Gliederung in den einzelnen Untertabellen nicht übereinstimmen. Beispielsweise könnten in der Tabelle der Abbildung 6.2 10 Summanden a_i zur Summe $\sum_1$, 15 b_i zur Summe $\sum_2$, usw. und nur 4 v_i zur letzten Zwischensumme $\sum_m$ beitragen. In solchen Fällen müssen die nicht "zusammenpassenden" Gliederungen durch leere Kategorien ergänzt werden. Ist die größte Anzahl der Kategorien (hier der Summanden) zu einer Zwischensumme in der Tabelle Abb. 6.2 15, so muss die erste Untertabelle (1. Zeile der Abbildung 6.3) mit 5 und die letzte mit 11 leeren Tabellenfeldern aufgefüllt werden. Dabei ist es für die Summation völlig unbedeutend, ob die leeren Kategorien (Dummy-Kategorien) jeweils an den Anfang, ans Ende oder zwischen die besetzten Kategorien gestreut werden, denn davon hängt nur die Summenbildung der neuen Gliederungen ab und diese Summenwerte werden niemals veröffentlicht.

Die Umstrukturierung einer n-dimensionalen Tabelle lässt sich ganz analog bewerkstelligen, indem man nach dem Muster einer eindimensionalen Tabelle ein Gliederungskriterium nach dem anderen umstellt und ergänzt, bis die resultierende Tabelle nicht mehr durch Zwischensummen untergliedert und daher als vollständig zu bezeichnen ist. Zur Begründung betrachte man die in der vorangestellten Abbildung aufgeführte, mehrfach durch Zwischensummen unterteilte Zeile als n-dimensionale Tabelle, deren Gliederung nach einem beliebig herausgegriffenen Ordnungskriterium zu dieser Zeile geführt hat. Alle Elemente der Zeile sind dann n-1-dimensionale Tabellen, deren einander entsprechende Werte zu addieren sind, so dass auch hier das Assoziativ- und das Kommutativgesetz zur Umstellung nach einem neuen Ordnungskriterium ausgenutzt werden kann. Diese Betrachtungen motivieren die

Definition:

Eine Statistiktabelle heißt vollständig, wenn die Addition von Tabellenwerten über jedes Gliederungskriterium (über jeden Index) immer zu genau einer Summe , der Randsumme, führt; die Gliederungskriterien, die zu keiner Zwischensumme Anlass geben, werden als elementare Gliederungen bezeichnet, ihre Indizes als Elementarindizes.

In diesem Sinne ist eine Untertabelle eine vollständige Tabelle, wenn man sie aus der Untertabellenhierarchie herausgelöst betrachtet. Die mehrfach durch Zwischensummen unterteilte Gesamttabelle hingegen ist nicht vollständig; sie muss durch Aufstocken der Dimension erst in eine vollständige Tabelle überführt werden, die dann keine Zwischensummen mehr hat.

Die (Elementar-)Dimension einer vollständigen Tabelle ergibt sich dann als Summe der höchsten Aggregationsstufen bezüglich jedes durch die ursprüngliche Tabelle gegebenen Gliederungskriteriums, vermindert um die Anzahl dieser Gliederungskriterien, wobei die unterste Aggregationsstufe gleich 1 gesetzt wurde.

Im Falle der zweidimensionalen Statistik des steuerbaren Umsatzes von 1994 mit einer Wirtschaftssystematik mit 7 Aggregationsstufen und der in NRW üblichen regionalen Gliederung mit 4 Aggregationsstufen, erhält man als vollständige Tabelle eine neundimensionale Tabelle.

6.2.2.1 Regeln zur Handhabung der durch Aufstockung hinzukommenden Werte

Die Aufstockung der Dimension führt bei realen Statistiktabellen in der Regel zu sehr umfangreichen, hochdimensionalen, vollständigen Tabellen, die gegenüber den ursprünglichen, mehrfach durch Zwischensummen unterteilten Tabellen durch Einfügen zusätzlicher Summen unterschiedlicher Aggregation, der „Sternchensummen“, und durch Eintragung strukturgebender Tabellenfelder, der Dummies, erweitert worden sind. Dabei kommt den meisten Dummy-Werten dieselbe Bedeutung zu wie den strukturellen Nullen: Sie können nicht zur Sicherung geheimer Werte gesperrt werden. Andere bei der Aufstockung zusätzlich einzutragende Werte, wie die „Sternchensummen“, sind als Sicherungspartner primär geheimer Werte besonders zu bevorzugen, weil sie in Veröffentlichungstabellen nicht auftreten, also wie bereits gesperrte, aber selbst nicht zu schützende Werte wirken.

Wenn diese Aufstockung von dem für die Wahrung der Geheimhaltung sensibler Daten verantwortlichen Fachstatistiker unter ausschließlicher Verwendung von Realdaten durchgeführt wird, oder wenn die zur Sicherung anstehenden Tabellendaten von vorne herein, d.h. nach fachlichen Gesichtspunkten bereits so strukturiert werden, dass nur noch vollständige Tabellen vorliegen, dann kann das im ersten Teil dieser Darstellung beschriebene Quaderverfahren ohne weitere Vorbereitungen angewendet werden; es bietet dann einen hinreichenden Schutz gegen zu genaues Rückrechnen primär geheimer Werte. Für die Sicherung der Tabellendaten ist dies zweifellos der Königsweg.

Ist aber die zur Bearbeitung mit dem Quaderverfahren vorgelegte Statistiktabelle noch durch Zwischensummen untergliedert und soll sie dem gemäß unmittelbar vor der Anwendung des Quaderverfahrens aufgestockt werden, so erfordert die Vielfalt der Aufstockungsmöglichkeiten insbesondere hinsichtlich der neuzubildenden Summen die Einrichtung von Platzhaltern als Tabellenwerte, über deren Inhalt meist nichts bekannt ist. Es kann nicht einfach die erste beste Umstrukturierung durchgeführt werden; der Fachstatistiker hätte u.U. eine ganz andere Aufstockung vorgenommen, womit dann auch ein ganz anderes Muster von Sekundärsperrungen entstanden wäre.

Durch Eintragung von Platzhaltern anstelle der real berechenbaren neuen Summen (Sternchensummen) und der strukturellen Dummy-Werte wird dieser Vielgestaltigkeit wenigstens zum Teil Rechnung getragen. Eine gewisse Willkür bleibt unvermeidbar, weil die Positionen der strukturellen Dummies für die Quaderauswahl wenigstens temporär festgelegt werden müssen. In der Möglichkeit, Dummypositionen zu verändern, verbirgt sich ein beträchtliches Optimierungspotential. Diese Idee wird in nachfolgendem Tabellenbeispiel angesprochen. Sie ist im EDV-Programm QUIT in Gestalt „deformierter vollständiger Quader“ realisiert.

Im Folgenden wird die Behandlung von Dummies und Sternchensummen bei der Quaderauswahl für den Fall positiver Tabellen ggf. mit Berücksichtigung von Schätzintervallen diskutiert. - Sind in der betrachteten Statistik positive und negative Tabellenwerte zu erwarten, kann Intervallschutz bei vollständigen Tabellen mit Schätzintervallangaben gewährleistet werden, oder man muss auf Intervallschutz verzichten.

Während die Positionen der Dummies bereits durch den Vorgang der Dimensionsaufstockung festgelegt werden, kann man bei positiven Tabellen auf die genaue Berechnung der Sternchensummen verzichten, weil sie weder veröffentlicht werden noch durch ihre tatsächlichen Werte einen Einfluss auf die Quaderauswahl haben: Weil Sternchensummen nicht veröffentlicht werden, sind sie wie bereits gesperrte Werte besonders zu bevorzugende Sicherungspartner und werden daher im Summenkriterium wie andere geheime Werte behandelt, deren tatsächlicher Wert nicht in der Quaderwertesumme erscheint (vergleiche Punkt 5.2).

Auch bei der Berechnung der Quaderspannweite (range) spielt der tatsächliche Wert von Sternchensummen keine Rolle, weil sie zur selben Quaderteilgesamtheit gehören wie die betreffenden Nachbarwerte im Tabelleninneren, aus denen sie hervorgehen. Sternchensummen sind bei den hier betrachteten positiven Tabellen daher stets größer oder höchstens genau so groß wie die zugehörigen Quaderwerte der gleichen Teilgesamtheit im Inneren der aufgestockten Tabelle, sodass ihr tatsächlicher Wert für die range-Berechnung keine Bedeutung hat. An dieser Aussage ändert sich auch nichts, wenn noch Schätzintervalle zu berücksichtigen sind, weil die Schätzintervalle der Sternchensummen die Schätzintervalle der zur selben Quaderteilgesamtheit gehörigen Werte im Tabelleninneren nicht einengen können. Die Schätzintervalle der Sternchensummen sind eben keinesfalls als kleiner anzusetzen wie die der realen Werte, die zu diesen Schätzintervallen beitragen.

Bei Vorliegen positiver Tabellen und ggf. auch bei zu berücksichtigenden Schätzintervallen ergibt sich aus den beiden letzten Absätzen folgende Regel für die Handhabung von Sternchensummen bei der sekundären Geheimhaltung:

Regel 1: Der Platzhalter eines durch die Aufstockung neu einzufügenden Summenwertes ist, wie jeder andere geheime Tabellenwert auch, ein bei der Quaderauswahl besonders zu bevorzugender Sicherungspartner, dessen Wert samt Schätzintervall bei der Spannweitenbestimmung aber unberücksichtigt bleibt und der selbst nicht vor Rückrechnung zu schützen ist.

Eine analoge Regel lässt sich auch für die Dummy-Felder herleiten. Wie oben bereits bemerkt, muss man dabei berücksichtigen, dass leere Dummy-Felder wie strukturelle Nullen nicht als Wert eines Sicherungsquaders in Frage kommen sollten. Andererseits lässt sich auf Dummy-Werte als Sicherungspartner nicht ganz verzichten, wie man bei Betrachtung von das fehlende Hinterland kreisfreier Städte ergänzenden Dummies leicht feststellt:

Dummies, die bei der Dimensionsaufstockung das „Hinterland“ einer kreisfreien Stadt auffüllen, können nicht alle einen Tabellenwert Null besitzen, weil ein nur mit Nullen besetztes Hinterland auch nur einen verschwindenden Wert für die kreisfreie Stadt ergäbe. Mit anderen Worten: Tragen ausschließlich Dummy-Werte zu von Null verschiedenen realen Summen bei, so können diese als Sicherungspartner geheimer Werte dienen. Dabei kann man für die Spannweitenberechnung auch annehmen, dass die Werte dieser Dummies und ihre Schätzintervalle genauso groß sind wie die der Randsummenwerte, zu denen sie beitragen, denn die Verteilung der Summenwerte auf ihr „Hinterland“ ist beliebig.

Umgekehrt lässt sich sagen, dass ein Dummy-Wert, der mit anderen realen Tabellenwerten zu einer realen Summe beiträgt, immer nur einer strukturellen Null entsprechen kann, die, wenn sie gesperrt würde, durch ihren Wegfall in der zu veröffentlichenden unaufgestockten Tabelle eine Geheimhaltungslücke hinterließe. Dummies, die bezüglich irgendeiner Gliederung zusammen mit realen Tabellenwerten zu einer realen Randsumme aufaddiert werden, müssen wie strukturelle Nullen offen bleiben. Werden sie zu einer Sternchensumme aufsummiert, so ist die Sternchensumme Quaderelement.

Die Handhabung von Dummy-Werten kann damit zu folgender Regel verdichtet werden:

Regel 2: Trägt ein Dummy bezüglich eines Gliederungskriteriums mit anderen real existierenden Tabellenwerten zu einer Summe bei, so ist er kein Sicherungspartner für einen geheimen Wert, der zugehörige Quader zu verwerfen; anderenfalls wirkt er wie ein nicht zu sichernder geheimer Wert, der bei der Spannweitenbestimmung unberücksichtigt bleibt.

Technische Anmerkungen

1. Um das bisher eingesetzte Quaderverfahren ohne weitere Modifikationen auch auf dimensionsaufgestockte Tabellen anwenden zu können und dabei o. g. Regeln 1 und 2 angemessen zu berücksichtigen, bietet sich die in Abschnitt 5.3 diskutierte externe Gewichtung als geeignetes Hilfsmittel an. Dabei reicht es für die Steuerung des Verfahrens aus, nur die Dummy-Werte und die durch die Aufstockung der Ta-

bellendimension neu entstandenen Sternchensummen geeignet bewertet und gewichtet in den Eingabe-Datenbestand einzutragen (im Folgenden werden keine Schätzintervalle berücksichtigt):

- Dummies, die nach Regel 2 wie strukturelle Nullen zu behandeln sind, werden im Eingabebestand als leere Tabellenfelder geführt. Sie sind dadurch als Sperrkandidaten ausgeschlossen.

- Durch Regel 2 nicht ausgeschlossene Dummies und Sternchensummen werden mit betrags- mäßig großen negativen Gewichten versehen um zu erreichen, dass sie – wie von Regel 1 und 2 verlangt – besonders bevorzugte Sperrkandidaten sind, die aber selbst nicht gesichert werden müssen.

- Sperrbare Dummies und Sternchensummen werden mit sehr großen positiven Tabellenwerten versehen (z.B. Eckfeldsummen), um zu erreichen, dass sie durch Auswahl der minimalen Werte der Quaderteilgesamtheiten niemals in die Spannweitenberechnung eingehen.

- Sperrbare Dummies und Sternchensummen werden als offene (nicht geheime) Tabellenwerte im dafür vorgesehenen Wertartfeld des Eingabebestandes markiert; sie sind gemäß Regel 1 und 2 nicht zu sichernde Werte.

2. Das oben angegebene Vorgehen bei der Sicherung einer aufgestockten Tabelle ist äußerst CPU-intensiv, weil sich die Suche eines Quaders zum Schutze eines geheimen Tabellenwertes nicht mehr auf eine kleine Untertabelle konzentriert, sondern weil stattdessen die gesamte und dazu auch noch durch die Aufstockung erheblich erweiterte Tabelle abgetastet werden muss. Hinzu kommt außerdem noch der exponentielle Zusammenhang der Anzahl elementarer Rechenoperationen mit der Tabellendimension, die durch die Dimensionsaufstockung beträchtlich zunimmt (zur Abschätzung der Rechenzeit vergleiche im ersten Teil den Punkt 2.2.2).

   Da andererseits eine Dimensionsaufstockung nur angezeigt ist, wenn Sperrungen in den Randsummen der Untertabellen auftreten und da Sperrungen in den Rand erfahrungsgemäß seltener vorkommen als im Inneren von Untertabellen (vergleiche Anhang A.1, graphische Darstellungen), bietet sich zumindest ein zweistufiges Vorgehen an, wonach im ersten Schritt alle primär geheimen Tabellenwerte auf unterstem Aggregationsniveau ohne Aufstockung der Dimension gesichert werden und wonach erst im zweiten Schritt die noch verbliebenen Sicherungen, die zu Randsperrungen führen, nach der Dimensionsaufstockung erfolgen. Nähere Ausführungen dazu finden sich im Abschnitt 6.2.2.3.

Die Gegenbeispieltabelle (Abbildung 6.1) wird nach Aufstockung zu einer vierdimensionalen Tabelle mit dem Quaderverfahren „ohne Intervallschutz“ mit einer Nullensperrung oder durch zwei Summensperrungen vollständig gesichert, je nachdem, ob Nullwerte als Sperrpartner zugelassen werden oder nicht. Bei dieser Tabelle genügt das Quaderverfahren „ohne Intervallschutz“, weil auch negative Tabellenwerte vorkommen können. Bei Verzicht auf Intervallschutz behalten die Regeln 1 und 2 - wie oben bemerkt - auch bei nicht positiven Tabellen ihre Gültigkeit.

Abb. 6.4

Behebung der Rückrechenbarkeit der Beispieltabelle in Abb. 6.1

X_1	0	X_1	X_2	0	X_2	X_3	0	X_3	10	*
X_4	0	X_4	0 ⊗	X_5	X_5	0	X_6	X_6	20	*
20	0	20	X_2	X_5	X_2+X_5	X_3	X_6	X_3+X_6	30	
0	0	0	X_7	0	X_7	X_8	0	X_8	40	
0	0	0	0	X_9	X_9	0	X_{10}	X_{10}	20	
0	0	0	X_7	X_9	X_7+X_9	X_8	X_{10}	X_8+X_{10}	60	
20	0	20	20	10	30	15	25	40	90	

⊗ wird als einziger Wert gesperrt, wenn Nullen sperrbar sind.

* werden keine Nullen akzeptiert, sperrt das Programm die beiden Randsummenwerte 10 und 20.

Wichtige Anmerkung:

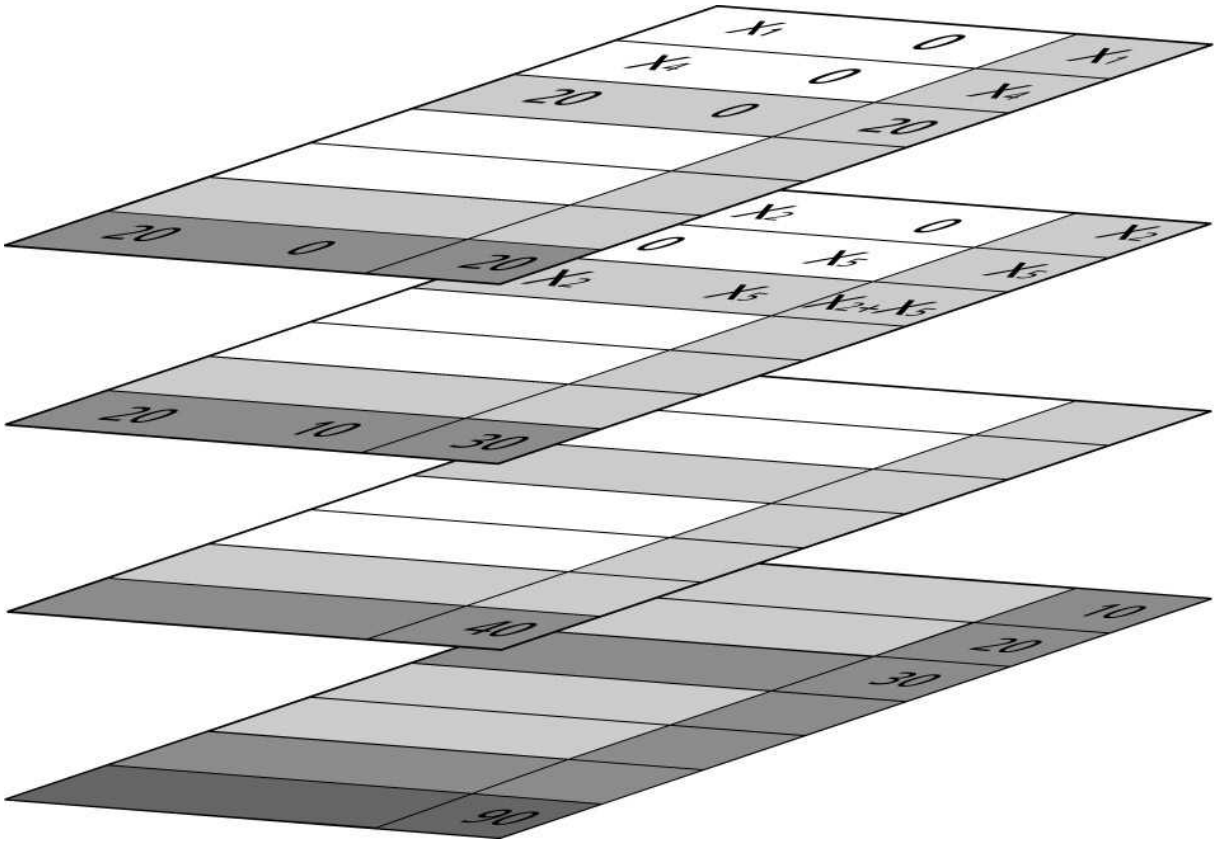
Wenn die Sperrung „⊗„ eingetragen ist, muss für den Summenwert X_2 eine neue Variable eingetragen werden, so dass die Bestimmungsgleichung $X_2 + X_7 = 20$ nicht mehr gilt!

Das Ergebnis der Quadersicherung in der vollständigen Tabelle Abb. 6.4 lässt sich anhand einer aufgestockten dreidimensionalen Tabelle veranschaulichen: Dazu ordnet man die drei Spalten-Streifen der zweidimensionalen Tabelle, Abbildung 6.1, aus den jeweils zwei zu einer Zwischensummenspalte beitragenden Spalten samt ihrer Zwischensummenspalte, gemäß Abb. 6.5 übereinander an:

Der erste, ganz linke Spalten-Streifen liegt zu oberst, der zweite, mittlere, darunter, gefolgt vom dritten ganz rechten Spalten-Streifen (ohne die Randsummenspalte dritter Aggregationsstufe). Die Randsummenspalte (dritte Aggregationsstufe) ist als Randsumme eines vierten, unter den anderen drei Streifen anzuordnenden Spalten-Streifens mit derselben Gliederungsstruktur aufzufassen. Die beiden anderen Spalten dieses vierten „Summenstreifens“ enthalten die Sternchensummen, die aus den darüber liegenden Tabellenwerten bezüglich der neuen dritten Tabellendimension zu berechnen sind.

Abb. 6.5

Aufstockung der Spaltengliederung der Gegenbeispieltabelle von Abb. 6.1 in schematisierter Darstellung



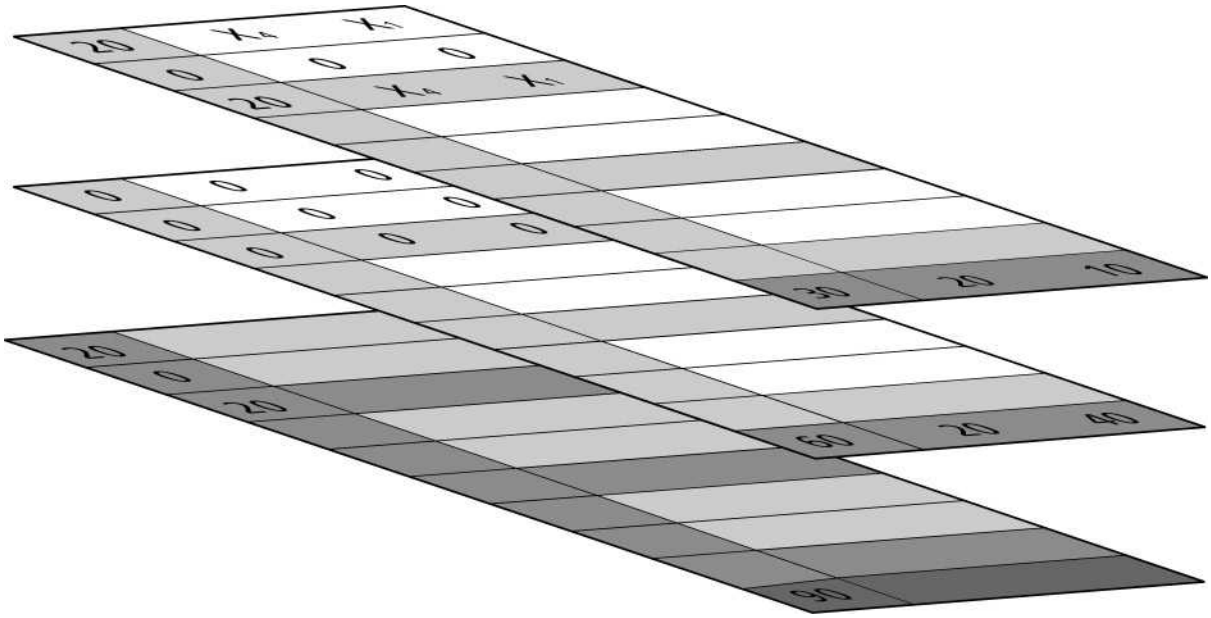
Die oberste zweidimensionale Tabelle dieser zur dreidimensionalen aufgestockten (aber noch nicht vollständigen) Tabelle enthält ein Karree von gesperrten Werten, das kein „Gegenstück“ in einer der darunter liegenden Streifen hat: Im zweiten und auch im dritten Streifen kann aber mit den drei bereits gesperrten Werten und einer Null, (X_2; X_2) in der ersten Zeile und (0; X_5) in der zweiten Zeile des zweiten Streifens bzw. (X_3; X_3) und (0; X_6) im dritten Streifen ein Karree zur Sicherung des obersten Karrees aufgebaut werden, wenn Nullwerte als Sperrpartner zugelassen sind. In diesem Fall wurde das Karree im zweiten Streifen gewählt - in Bezug auf die Quaderauswahlkriterien ist es zu dem des dritten Streifens völlig gleichwertig.

Werden aber die in der Tabelle Abb. 6.1 vorkommenden Nullen als strukturelle Nullen aufgefasst und daher als Sicherungspartner ausgenommen, so lässt sich ein Karree zur Sicherung des obersten Karrees nur noch mit den Sternchensummen der ersten Spalte des vierten (untersten) Streifens realisieren. Dazu muss man aber auch die beiden ersten Zeilenwerte der dritten Spalte des vierten Streifens als Gegenstück zur dritten Spalte des ersten Streifens sperren, d.h. die beiden Randsummenwerte dritter Aggregationsstufe müssen gesperrt werden, in Übereinstimmung mit dem EDV-Verfahren bei Aufstockung zur vollständigen vierdimensionalen Tabelle.

Da hier scheinbar schon eine zur dreidimensionalen Tabelle aufgestockte Tabelle anstelle der vollständigen vierdimensionalen ausreicht, um die gegebene mehrfach durch Zwischensummen unterteilte zweidimensionale Tabelle vollständig zu sichern, könnte man vermuten, dass das Quaderverfahren (oder ein anderes Sekundärsperrverfahren) auch bei Untertabellenabgleich einen hinreichenden Schutz bieten kann, wenn höchstens eine einzige Gliederung (mehrfach) durch Zwischensummen unterteilt ist und die anderen Gliederungen keine Zwischensummen enthalten. Dass dies nicht so ist, zeigt bereits die obige zweidimensionale Tabelle (Abb. 6.1), wenn man nicht die Spalten-, sondern die Zeilengliederung aufstockt:

Abb. 6.6

Aufstockung der Zeilengliederung der Gegenbeispieltabelle von Abb. 6.1 in schematisierter Darstellung



Bei der Aufstockung der Zeilengliederung der Tabelle, Abb. 6.1, kann man als obersten Zeilenstreifen die drei ersten Zeilen nehmen, als den zweiten darunter liegenden Zeilenstreifen die vierte bis sechste Zeile; darunter liegt dann der Summenstreifen. Er hat dieselbe Gliederungsstruktur wie die beiden darüber angeordneten Zeilenstreifen mit Werten, die sich bezüglich der neuen dritten Gliederung als Summe aus den darüber liegenden Werten ergeben. Dazu gehört auch die unterste Zeile der ursprünglichen Tabelle (Abb. 6.1) als Summenzeile der in der aufgestockten Tabelle darüber liegenden Summenzeilen zweiter Aggregationsstufe (siehe Abb. 6.6).

Die beiden oberen Streifen stimmen in ihren Sperrmustern bis auf die erste und dritte Spalte völlig überein. Aber auch die in der ersten und dritten Spalte des obersten Zeilenstreifens markierten geheimen Werte haben ihr „Gegenstück“ und zwar im untersten Summenstreifen, weil ja Sternchensummen nicht veröffentlicht werden und daher wie geheime Werte wirken (die selbst nicht zu sichern sind). Mit der so aufgestockten Tabelle hat man also eine dreidimensionale Tabelle mit nur einem mehrfach durch Zwischensummen unterteilten Gliederungsmerkmal gefunden, die mit dem Quaderverfahren und Untertabellenabgleich gesichert ist, die aber trotzdem noch die berechenbaren geheimen Werte X_1 und X_4 enthält.

Die nur bezüglich der Zeilengliederung aufgestockte dreidimensionale Tabelle stellt also ein Gegenbeispiel zu obiger Vermutung dar, dass n-dimensionale Tabellen mit nur einer durch Zwischensummen unterteilten Gliederung durch Untertabellenabgleich hinreichend gesichert werden könnten! Soll also ein hinreichender Quaderschutz gewährleistet sein, wird man im Allgemeinen nicht auf die Aufstockung der gegebenen zur vollständigen Tabelle verzichten dürfen (es sei denn, der betreffende Sicherungsquader enthält keine Zwischensummenwerte der so aufgestockten unvollständigen Tabelle; vergleiche den Abschnitt 6.2.2.3).

Solch eine vollständige Tabelle unterhält nun keine Wechselbeziehungen mit anderen Untertabellen der Gesamttabelle mehr, um derentwegen sie bezüglich irgendwelcher Summensperrungen abgeglichen werden müsste; sie kann daher mit dem Quaderverfahren hinreichend gesichert werden.

Es bleibt noch die Frage, wie die Gliederungsmerkmale der neuen Gliederung (d.h. der Gliederung der aufgestockten Tabelle nach Elementarindizes) dargestellt werden sollen. Die Antwort darauf gibt bereits das Quaderverfahren mit Untertabellenabgleich: Die neu hinzukommenden Gliederungsmerkmale sind im alten Tabellenabgleichsverfahren schon vorhanden: Es sind die Positionsindizes, die die geometrische Lage der Untertabellen in der Gesamttabelle festlegen (vgl. Abschnitt 1.2.1 und insbesondere Abb. 1.8). Die Gesamtheit der Gliederungskriterien der aufgestockten Tabelle umfasst demnach die alten vorgegebenen Nutzerindizes unterster Aggregation mit den in den jeweiligen Gliederungen meisten Ausprägungen nebst Randsummenindizes (anschließende Indizes zur um 1 höheren Aggregationsstufe) und die Positionsindizes zu jedem alten Gliederungskriterium und zu jeder Aggregationsstufe bis zur zweithöchsten Aggregation samt den anschließenden Positionsindizes zur um 1 höheren Aggregation der Randsummentabellen.

Lediglich die in die aufgestockte Tabelle einzufügenden Dummies haben in dieser Indexmenge noch keine Entsprechung. Dummy-Werte sind aber für die Quaderauswahl nur von qualitativer Bedeutung: Man muss unterscheiden können zwischen Dummies, die als Sicherungspartner eines geheimen Wertes in Betracht kommen und den anderen, denen nur strukturelle Bedeutung zukommt. Diese Unterscheidungsmöglichkeit ist aber bereits durch die vorhandenen Indizes gegeben, denn daraus werden auch die zur Bewertung der Dummies aufgestellten Regeln 1 und 2 hergeleitet (vgl. 7.1.2, Abb. 7.3 mit anschließend aufgeführten Dummy-Beispielen).

Es liegt daher nahe, die aufgestockte Tabelle als Fiktion zu begreifen, die für die Auswahl eines hochdimensionalen Sicherungsquaders nur als Hilfe zur Auffindung der u.U. in mehreren Untertabellen der gegebenen Statistiktabelle liegenden Quaderwerte dient. Im Folgenden wird zunächst von einer auf Datenträger real zu erstellenden vollständigen Tabelle ausgegangen; der Vorteil gegenüber einer fiktiven Tabelle liegt darin, dass Dummies nur einmal bewertet werden müssen und nicht bei jedem Aufbau der vielen Sicherungsquader, an denen sie beteiligt sein können.

6.2.2.2 Aufstockung der Beispieltabelle

Die Abbildung 6.7 gibt die Verteilung der Sperrungen in der schon im ersten Abschnitt verwendeten zweidimensionalen Beispieltabelle wieder, wie sie bei Aufstockung zur vollständigen vierdimensionalen Tabelle entsteht. Um die Diskussion auf die Wirkung der Aufstockung zu beschränken, wurden dabei keine Nullwerte verwendet, keine Randschranken gesetzt und es wurde auf Intervallschutz verzichtet. Das erhaltene Sperrmuster ist daher mit dem der Tabelle Abb. 1.7 des ersten Abschnitts zu vergleichen, die unter analogen Bedingungen gesichert worden ist, jedoch nur als zweidimensionale durch Zwischensummen untergliederte Tabelle.

Wie unter Punkt 6.2.2.1, technische Anmerkungen, bereits ausgeführt, verlangen die bei der Vervollständigung einzufügenden Dummy- bzw. Sternchensummenwerte besondere Vorkehrungen bei der Steuerung des bisher nur auf Tabellen mit Zwischensummen angewendeten EDV-Programms: Um zu erreichen, dass als Sperrkandidaten zugelassene neu eingefügte Tabellenwerte bei der Quaderauswahl gemäß den Regeln 1 und 2 ebenso bevorzugt werden wie bereits gesperrte Werte, bietet sich eine Gewichtung dieser neuen Tabellenwerte mit negativen Zahlen an. Darüber hinaus sollen Dummies und Sternchensummen nicht zur Quaderwertespannweite beitragen, wenn denn ein Intervallschutz gewünscht wird. Das kann man erreichen, indem man diese Werte durch sehr große Tabellenwerte ersetzt, weil bei der diesbezüglichen Auswahl des jeweils kleinsten Wertes einer Quaderteilgesamtheit die großen Tabellenwerte ausgesondert werden.

Die Bearbeitung einer vollständigen Tabelle mit dem Quaderverfahren ist dem gemäß auch ein Anwendungsbeispiel für die externe Gewichtung, die in Abschnitt 5.3 besprochen wurde. Und zwar handelt es sich hier um eine von der Position der Tabellenfelder abhängige Gewichtung gemäß 5.3.1 Unterpunkt c). Die besondere Bevorzugung von sperrbaren Dummies und Sternchensummen (letztere werden auch synonym als Dummies bezeichnet) wird in der Tabelle der Abbildung 6.8 durch besonders große Werte und betragsmäßig nicht sehr große negative Gewichte erzwungen. Bei der Festlegung dieser Werte sollte die hinter c) angeführte technische Anmerkung nicht außer Acht bleiben. Die genaue Festlegung von Dummywerten und ihren Gewichten wird im Folgenden noch erarbeitet.

Die Abbildung 6.7 umfasst nur die in der Veröffentlichungstabelle aufzuführenden Tabellenfelder; die durch die Dimensionsaufstockung erzwungene wesentliche Erweiterung bleibt dabei verborgen. Ein Vergleich der beiden Abbildungen 1.7 und 6.7 zeigt, dass durch die Vervollständigung der Veröffentlichungstabelle eine gewisse Umstrukturierung des Sperrmusters auftritt, die nicht mehr durch die zu veröffentlichenden Werte und deren Positionen innerhalb der Tabelle allein zu erklären ist; es muss auch die geometrische Anordnung der als Sperrpartner zugelassenen Dummy-Werte und Sternchensummen berücksichtigt werden. Dazu kann man die gegebene zweidimensionale Tabelle als aufgestockte vierdimensionale Tabelle recht übersichtlich als ebenes Zahlentableau darstellen: Abbildung 6.8 zeigt die vollständige Beispieltabelle in Matrixform.

Aufstockung der Beispieltabelle

Abb. 6.7

1. Schlüssel	2. Schlüssel														
	ACD	ACC	ACB	ACA	AC	ABC	ABB	ABA	AB	AAD	AAC	AAB	AAA	AA	A
00000134	112 5 S	10 2 P	1.445 20	549 12	2.116 39	4.128 34	345 15	211 12	4.684 61	321 21 S	0 0	0 0	95 2 P	416 23	7.216 123
00000133	40 1 P	66 4 S	0 0	23 3	129 8	2.567 44	2.332 30	432 21	5.331 95	732 51	644 34	0 0	0 0	1.376 85	6.836 188
00000132	723 9	254 11	327 5	543 19	1.847 44	1.123 64	4.427 59	1.632 26	7.182 149	432 23	0 0	234 36	0 0	666 59	9.695 252
00000131	2.156 33	1.342 23	1.111 17	99 4	4.708 77	590 11	2.334 28	342 9	3.266 48	34 3 S	0 0	0 0	256 17 S	290 20	8.264 145
00000130	3.031 48 S	1.672 40 S	2.883 42	1.214 38	8.800 168	8.408 153	9.438 132	2.617 68	20.463 353	1.519 98	644 34	234 36	351 19	2.748 187	32.011 708
00000125	321 5 S	11 3 S	411 18	0 0	743 26	0 0	56 5 S	0 0	56 5 S	712 50	3.421 84	0 0	0 0	4.133 134	4.932 165 S
00000124	56 4 S	12 1 P	2.152 29	399 11	2.619 45	0 0	123 10	0 0	123 10	345 44	2.612 61	55 3	0 0	3.012 108	5.754 163
00000123	99 8	311 10	754 19	345 16	1.509 53	221 7	34 2 P	73 6	328 15 S	123 23	321 41	567 32	43 4	1.054 100	2.891 168 S
00000122	1.837 33 S	19 1 P	88 4	0 0	1.944 38	0 0	621 13	0 0	621 13	1.015 89	2.221 52	96 18	641 8	3.973 167	6.538 218
00000121	344 15	298 13	0 0	934 9	1.576 37	0 0	74 8	0 0	74 8	0 0	231 33	0 0	875 72	1.106 105	2.756 150
00000120	2.657 65	651 28	3.405 70	1.678 36	8.391 199	221 7	908 38	73 6	1.202 51	2.195 206	8.806 271	718 53	1.559 84	13.278 614	22.871 864
00000113	53 2 P	221 8 S	29 3	1.001 19	1.304 32	0 0	0 0	0 0	0 0	11 2 P	0 0	21 2 P	0 0	32 4	1.336 36
00000112	423 18	0 0	0 0	0 0	423 18	0 0	261 5 S	34 2 P	295 7	745 71 S	0 0	67 8 S	0 0	812 79	1.530 104
00000111	28 5	0 0	0 0	0 0	28 5	0 0	0 0	0 0	0 0	148 25	0 0	81 7	0 0	229 32	257 37
00000110	504 25 S	221 8 S	29 3	1.001 19	1.755 55	0 0	261 5 S	34 2 P	295 7	904 98	0 0	169 17	0 0	1.073 115	3.123 177
00000100	6.192 138	2.544 76	6.317 115	3.893 93	18.946 422	8.629 160	10.607 175 S	2.724 76 S	21.960 411	4.618 402	9.450 305	1.121 106	1.910 103	17.099 916	58.005 1.749

Legende:

Wert
Berichtspfl.

10.000
100 P

Sperrvermerk (P=primär, S=sekundär)

1. Schlüssel	2. Schlüssel																			
	ACD	ACC	ACB	ACA	AC	ABC	ABB	ABA	ABA'	AB	AAD	AAC	AAB	AAA	AA	AAD'	AAC'	AAB'	AAA'	A
00000134	112 5 S	10 2 P	1.445 20	549 12	2.116 39	4.128 34	345 15	211 12	0 0 D	4.684 61	321 21 S	0 0	0 0	95 2 P	416 23	60.000 4 GD	60.000 4 GD	60.000 4 SD	60.000 4 GD	7.216 123
00000133	40 1 P	66 4 S	0 0	23 3	129 8	2.567 44	2.332 30	432 21	0 0 D	5.331 95	732 51	644 34	0 0	0 0	1.376 85	60.000 4 GD	60.000 4 GD	60.000 4 SD	60.000 4 SD	6.836 188
00000132	723 9	254 11	327 5	543 19	1.847 44	1.123 64	4.427 59	1.632 26	0 0 D	7.182 149	432 23	0 0	234 36	0 0	666 59	60.000 4 SD	60.000 4 SD	60.000 4 SD	60.000 4 SD	9.695 252
00000131	2.156 33	1.342 23	1.111 17	99 4	4.708 77	590 11	2.334 28	342 9	0 0 D	3.266 48	34 3 S	0 0	0 0	256 17 S	290 20	60.000 4 GD	60.000 4 SD	60.000 4 SD	60.000 4 GD	8.264 145
00000131'	0 0 D	0 0 D	0 0 D	0 0 D	0 0 D	0 0 D	0 0 D	0 0 D	0 4 SD	0 0 D	0 0 D	0 0 D	0 0 D	0 0 D	0 0 D	60.000 4 SD	60.000 4 SD	60.000 4 SD	60.000 4 SD	0 0 D
00000130	3.031 48 S	1.672 40 S	2.883 42	1.214 38	8.800 168	8.408 153	9.438 132	2.617 68	0 0 D	20.463 353	1.519 98	644 34	234 36	351 19	2.748 187	60.000 4 GD	60.000 4 GD	60.000 4 SD	60.000 4 SD	32.011 708
00000125	321 5 S	11 3 S	411 18	0 0	743 26	0 0	56 5 S	0 0	0 0 D	56 5 S	712 50	3.421 84	0 0	0 0	4.133 134	60.000 4 GD	60.000 4 GD	60.000 4 SD	60.000 4 SD	4.932 165 S
00000124	56 4 S	12 1 P	2.152 29	399 11	2.619 45	0 0	123 10	0 0	0 0 D	123 10	345 44	2.612 61	55 3	0 0	3.012 108	60.000 4 GD	60.000 4 GD	60.000 4 SD	60.000 4 SD	5.754 163
00000123	99 8	311 10	754 19	345 16	1.509 53	221 7	34 2 P	73 6	0 0 D	328 15 S	123 23	321 41	567 32	43 4	1.054 100	60.000 4 SD	60.000 4 GD	60.000 4 SD	60.000 4 SD	2.891 168 S
00000122	1.837 33 S	19 1 P	88 4	0 0	1.944 38	0 0	621 13	0 0	0 0 D	621 13	1.015 89	2.221 52	96 18	641 8	3.973 167	60.000 4 GD	60.000 4 GD	60.000 4 SD	60.000 4 SD	6.538 218
00000121	344 15	298 13	0 0	934 9	1.576 37	0 0	74 8	0 0	0 0 D	74 8	0 0	231 33	0 0	875 72	1.106 105	60.000 4 SD	60.000 4 SD	60.000 4 SD	60.000 4 SD	2.756 150
00000120	2.657 65	651 28	3.405 70	1.678 36	8.391 199	221 7	908 38	73 6	0 0 D	1.202 51	2.195 206	8.806 271	718 53	1.559 84	13.278 614	60.000 4 SD	60.000 4 SD	60.000 4 SD	60.000 4 SD	22.871 864
00000113	53 2 P	221 8 S	29 3	1.001 19	1.304 32	0 0	0 0	0 0	0 0 D	0 0	11 2 P	0 0	21 2 P	0 0	32 4	60.000 4 GD	60.000 4 GD	60.000 4 GD	60.000 4 SD	1.336 36
00000112	423 18	0 0	0 0	0 0	423 18	0 0	261 5 S	34 2 P	0 0 D	295 7	745 71 S	0 0	67 8 S	0 0	812 79	60.000 4 GD	60.000 4 GD	60.000 4 GD	60.000 4 SD	1.530 104
00000111	28 5	0 0	0 0	0 0	28 5	0 0	0 0	0 0	0 0 D	0 0	148 25	0 0	81 7	0 0	229 32	60.000 4 SD	60.000 4 SD	60.000 4 SD	60.000 4 SD	257 37
00000111'	0 0 D	0 0 D	0 0 D	0 0 D	0 0 D	0 0 D	0 0 D	0 0 D	0 4 SD	0 0 D	0 0 D	0 0 D	0 0 D	0 0 D	0 0 D	60.000 4 SD	60.000 4 SD	60.000 4 SD	60.000 4 SD	0 0 D
00000111''	0 0 D	0 0 D	0 0 D	0 0 D	0 0 D	0 0 D	0 0 D	0 0 D	0 4 SD	0 0 D	0 0 D	0 0 D	0 0 D	0 0 D	0 0 D	60.000 4 SD	60.000 4 SD	60.000 4 SD	60.000 4 SD	0 0 D
00000110	504 25 S	221 8 S	29 3	1.001 19	1.755 55	0 0	261 5 S	34 2 P	0 0 D	295 7	904 98	0 0	169 17	0 0	1.073 115	60.000 4 GD	60.000 4 GD	60.000 4 GD	60.000 4 SD	3.123 177
00000113'	60.000 4 GD	60.000 4 GD	60.000 4 SD	60.000 4 SD	60.000 4 SD	60.000 4 SD	60.000 4 GD	60.000 4 SD	60.000 4 SD	60.000 4 GD	60.000 4 GD	60.000 4 SD	60.000 4 GD	60.000 4 GD	60.000 4 SD	60.000 4 GD	60.000 4 GD	60.000 4 GD	60.000 4 GD	60.000 4 GD
00000112'	60.000 4 GD	60.000 4 GD	60.000 4 SD	60.000 4 SD	60.000 4 SD	60.000 4 SD	60.000 4 GD	60.000 4 GD	60.000 4 SD	60.000 4 SD	60.000 4 GD	60.000 4 SD	60.000 4 GD	60.000 4 SD	60.000 4 SD	60.000 4 GD	60.000 4 GD	60.000 4 GD	60.000 4 SD	60.000 4 SD
00000111'	60.000 4 SD	60.000 4 SD	60.000 4 SD	60.000 4 SD	60.000 4 SD	60.000 4 SD	60.000 4 GD	60.000 4 SD	60.000 4 SD	60.000 4 GD	60.000 4 SD	60.000 4 SD	60.000 4 SD	60.000 4 SD	60.000 4 SD	60.000 4 SD	60.000 4 GD	60.000 4 SD	60.000 4 SD	60.000 4 GD
00000111'	60.000 4 GD	60.000 4 GD	60.000 4 SD	60.000 4 SD	60.000 4 SD	60.000 4 SD	60.000 4 SD	60.000 4 SD	60.000 4 SD	60.000 4 SD	60.000 4 GD	60.000 4 SD	60.000 4 SD	60.000 4 GD	60.000 4 SD	60.000 4 GD	60.000 4 GD	60.000 4 SD	60.000 4 GD	60.000 4 SD
00000111''	60.000 4 SD	60.000 4 SD	60.000 4 SD	60.000 4 SD	60.000 4 SD	60.000 4 SD	60.000 4 SD	60.000 4 SD	60.000 4 SD	60.000 4 SD	60.000 4 SD	60.000 4 SD	60.000 4 SD	60.000 4 SD	60.000 4 SD	60.000 4 SD	60.000 4 SD	60.000 4 SD	60.000 4 SD	60.000 4 SD
00000100	6.192 138	2.544 76	6.317 115	3.893 93	18.946 422	8.629 160	10.607 175 S	2.724 76 S	0 0 D	21.960 411	4.618 402	9.450 305	1.121 106	1.910 103	17.099 916	60.000 4 SD	60.000 4 GD	60.000 4 GD	60.000 4 SD	58.005 1.749

Legende: Wert 10.000 / Berichtspfl. 100 P — Kennzeichnung (P=primäre Sperrung, S=sekundär Sperrung, SD=sperrbare Dummies, D=nicht sperrbare Dummies, GD=gesperrte Dummies)

Für die EDV-Programmsteuerung sind folgende Parameterwerte eingefügt worden: Der größtmögliche Tabellenwert; er wurde mit 61.000 veranschlagt; der sperrbare Dummy- bzw. Sternchensummenwert wurde als 60.000 angenommen; die Gewichtung der neu eingefügten Tabellenwerte erfolgte mit dem Faktor - 15; für den minimalen Tabellenwert ist 1 angesetzt worden. Die Abbildung 6.8 zeigt die durch einen Spalten- und einen Zeilenstreifen geränderte Beispieltabelle der Abbildung 6.7, wobei die Doppelsummenspalte und die Doppelsummenzeile (beide in dunkelster Schraffur) das Gesamttableau von rechts und unten umranden. Der ganz rechte Spalten- und der unterste Zeilenstreifen nehmen die Sternchensummen auf. Ihr gemeinsamer Wert ist, wie oben verfügt, 60.000; als Fallzahl wurde willkürlich 4 eingetragen. Die Sternchensummenfelder mit GD-Vermerken markieren die Werte vierdimensionaler Quader, die zur Sicherung nur der primär geheimen Tabellenwerte der gegebenen Beispieltabelle ausgewählt worden sind. Eine darüber hinausgehende Sicherung der Sekundärsperrungen ist bei der Bearbeitung einer vollständigen Tabelle nicht erforderlich, weil aufgrund der Vervollständigung nur eine einzige „Untertabelle“ existiert.

Außer einer Erweiterung der Gesamttabelle durch die Sternchensummentabellen in den Randstreifen ist noch eine Ergänzung des zweiten Spaltenstreifens durch eine Spalte mit Dummywerten auf insgesamt 4 Spalten erforderlich. Dadurch erhält der zweite Streifen dieselbe Spaltenanzahl wie die anderen Spaltenstreifen. Des Weiteren muss der erste Zeilenstreifen durch eine und der dritte Zeilenstreifen durch zwei Zeilen von Dummywerten ergänzt werden, um die gleiche Zeilenanzahl wie beim zweiten Zeilenstreifen zu erreichen. Alle Dummywerte dieser ergänzten Zeilen - und Spaltenstreifen, soweit sie gemeinsam mit anderen real vorhandenen Tabellenwerten aufaddiert werden können, sind gemäß Regel 2 als Werte von Sicherungsquadern tabu. Dies trifft nicht zu für die drei Tabellenfelder auf den Schnittstellen der mit Dummywerten versehenen Zeilen und Spalten, die daher als bevorzugte Sperrkandidaten mit demselben Tabellenwert (und derselben fiktiven Fallzahl) wie die Sternchensummen versehen wurden. Sie haben aber trotzdem keine Bedeutung für die Sicherung primär geheimer Werte, weil in der betreffenden Zeile und Spalte sonst nur tabuisierte Dummies als strukturelle Nullen eingetragen sind.

Erst nach dieser Erweiterung der schmaleren Zeilen- und Spaltenstreifen zur vollständigen Zeilen- und Spaltenanzahl der breitesten Streifen lassen sich die durch die Summenzeilen 130, 120, 110 und 100 und durch die Summenspalten AC, AB, AA und A abgegrenzten Untertabellen z.B. zeilenstreifenweise übereinander zu 3 dreidimensionalen Tabellen anordnen, deren Werte dann schließlich als Summe über diese dreidimensionalen Teiltabellen, einer vierdimensionalen Tabelle, die dreidimensionale Summenzeilen-Tabelle ergeben. Mit dieser geometrischen Deutung lässt sich das erhaltene Sperrmuster anschaulich erklären.

Zur Beschriftung der aufgestockten Tabelle, Abb. 6.8, ist anzumerken, dass durch die Einführung neuer Gliederungskriterien die alten ihren Sinn verlieren. Um aber einen direkten Bezug zur „Veröffentlichungstabelle“, Abb. 6.7 bzw. Abb. 1.7, herzustellen, wurden in Abb. 6.8 die ursprünglichen Gliederungsmerkmale eingetragen und die Ergänzungen, eingefügte Zeilen oder Spalten, durch analoge Gliederungsausprägungen, jedoch durchgestrichen, markiert.

Die Position der einzufügenden Zeile oder Spalte innerhalb des betreffenden Streifens ist beliebig, die Veröffentlichungstabelle bleibt davon unbeeinflusst. Anders verhält es sich mit dem Sperrmuster, weil von der Auswahl der

Einfügungspositionen die geometrische Anordnung der Primärsperrungen und damit auch die Quaderauswahl beeinflusst wird. Dadurch bietet sich u.U. eine zusätzliche Optimierungsmöglichkeit zur Verringerung der Anzahl von Sekundärsperrungen an. In der vorliegenden Beispieltabelle ließe sich die Verteilung der Primärsperrungen günstig beeinflussen, wenn man eine der Dummyzeilen ~~111~~ zwischen die Zeilen 112 und 113 einordnete. Dann lägen die Zeilen 112 und 123 mit Primärsperrungen in beiden Zeilen innerhalb der zugehörigen Untertabellen des zweiten Spaltenstreifens in gleicher Zeilenposition und ließen sich so in einem vierdimensionalen Quader zusammenfassen.

Die EDV-mäßige Bearbeitung der aufgestockten Tabelle erfolgt hier spaltenweise. Demnach findet das Programm den ersten zu sichernden primär geheimen Wert im Tabellenfeld (133; ACD). Er wird durch den vierdimensionalen $2^4 = 16$ Tabellenwerte umfassenden Quader {(134; ACD), (134; ACC), (133; ACD), (133; ACC), (~~113;~~ ACD), (~~113~~; ACC), (~~112~~; ACD), (~~112~~; ACC), (134; ~~AAD~~), (134; ~~AAC~~), (133; ~~AAD~~), (133; ~~AAC~~), (~~113~~; ~~AAD~~), (~~113~~; ~~AAC~~), (~~112~~; ~~AAD~~), (~~112~~; ~~AAC~~)} gesichert. - Jeder dieser Quaderwerte ist als Element einer vierdimensionalen Tabelle durch vier Indizes (4 Gliederungsmerkmalsausprägungen) geometrisch fixiert. - Dass hier nur zwei Indizes pro Quaderwert genügten, liegt an der in Abb. 6.8 verwendeten Matrixdarstellung der vierdimensionalen Tabelle begründet. –

Der betrachtete vierdimensionale Sicherungsquader deckt dem gemäß nur ein Karree von realen Tabellenwerten auf unterstem Aggregationsniveau ab. Die drei anderen Karrees mit den restlichen 12 geheimen Werten liegen als Projektionen des „realen Karrees“ in den Tabellen der mit Sternchensummen gefüllten Randstreifen sowie in der Schnitttabelle beider Randstreifen. Sie schaden der Veröffentlichungstabelle daher in keiner Weise! Die Sicherung eines primär geheimen Wertes auf unterstem Aggregationsniveau mit realen Quaderwerten, die sich ebenfalls alle auf unterstem Niveau befinden, führt zu genau denselben Sperrungen realer Tabellenwerte wie die Sicherung mit dem zugehörigen hochdimensionalen Quader der aufgestockten vollständigen Tabelle (vergleiche auch den 2. Punkt der „technischen Anmerkungen“ zum Unterpunkt 6.2.2.1), weil alle nicht zur ursprünglichen (unvollständigen) Tabelle gehörigen Quaderteile durch Projektion des realen Quaders ins Innere der Sternchensummentabellen entstehen und somit keine realen Sperrungen hervorbringen. Aus dieser Quadereigenschaft ergibt sich die in den technischen Anmerkungen angegebene Reduktionsmöglichkeit der CPU-Rechenzeit!

Den nächsten zu schützenden primär geheimen Tabellenwert findet das Programm beim spaltenweisen Vorgehen im Feld (113; ACD). Dieser Wert ist nicht mit einem Karree im Inneren der realen Untertabelle zu sichern; das Programm muss auf Randsummenwerte - nach Abb. 6.7 und 6.8 auf Werte in Zeile 110 - zurückgreifen. Beim Aufbau des entsprechenden vierdimensionalen Quaders können bereits primär und sekundär gesperrte Werte verwendet werden: Das Programm bildet im ersten Spaltenstreifen einen dreidimensionalen Quader mit den geheimen Werten der Zeile 134, den noch zu sperrenden Summenwerten in den Zeilen 130 und 110 mit den Spalten ACD, ACC sowie mit dem zu sperrenden Wert im Feld (113; ACC) und dem Pivot in derselben Zeile. Diesen dreidimensionalen Quader projiziert es in die Sternchensummenspalten ~~AAD~~ und ~~AAC~~ (in dieselben Zeilen). Da die Summenspalte AC von diesen Sperrungen nicht betroffen ist, kann die gesamte Sicherung des Pivots (113; ACD) vollständig im Inneren des ersten Spaltenstreifens erfolgen, ohne die beiden anderen Spaltenstreifen der realen Tabelle zu behelligen. Das gleiche gilt auch für die anderen noch zu sichernden primär geheimen Werte dieses Spaltenstreifens.

Man sieht: Zwischensummen ohne Sperreintragungen wirken wie Barrieren gegen Übertragungen von Sekundärsperrungen in andere durch die sperrungsfreie Zwischensumme abgetrennte Tabellenteile, weil die in solchen Fällen vorzunehmende Projektion des gesamten von der Sicherung betroffenen Tabellenteils ausschließlich ins Innere des zugehörigen Teils der Sternchensummen erfolgt bzw. in Sternchenrandsummen, die auch nicht in der Veröffentlichungstabelle erscheinen. Auf diesen Sternchensummenteil kann bei der Bearbeitung des abgetrennten Tabellenteils direkt verzichtet werden. Es genügt also, den durch sperrungsfreie Zwischensummen abgetrennten Tabellenteil für sich allein zu bearbeiten, wodurch sich der Umfang und insbesondere die Dimension des nach Sicherungsquadern abzusuchenden Tabellenteils entsprechend reduziert.

Eine einfache algebraische Erklärung für die Trennwirkung von sperrungsfreien Summen ist durch das Gleichungssystem zur Berechnung der geheimen Werte als Unbekannte gegeben: Die Unbekannten dieses Gleichungssystems tragen jeweils nur zu ihren sperrungsfreien Summen bei oder zu Zwischensummen, aus denen diese bestehen, und nicht zu Summen oder Zwischensummen eines anderen durch die sperrungsfreie Summentabelle abgetrennten Tabellenteils; sie kommen also nur in demjenigen Teilsystem von Bestimmungsgleichungen vor, das durch die sperrungsfreien Summen vom Gesamtgleichungssystem der Tabelle abgegrenzt wird. - Auf diese Möglichkeit der Unterteilung der Gesamttabelle hat der Autor bereits in seinem Papier zum internationalen Seminar zur statistischen Geheimhaltung 1994 in Luxemburg hingewiesen.

In der obigen Beispieltabelle bietet sich also an, den von der restlichen Tabelle abgetrennten Spaltenstreifen als dreidimensionale vollständige Tabelle zu behandeln. Da man die tatsächliche Verteilung der Sekundärsperrungen zum Zeitpunkt der Programmausführung nicht kennt, wird man zunächst Probeläufe mit kleineren Tabellenteilen durchführen. Auch dieses Vorgehen muss bei der Suche nach Einsparmöglichkeiten von Rechenzeit in Betracht gezogen werden.

Der Vergleich des ersten Spaltenstreifens der Veröffentlichungstabelle 6.7 mit der „zweidimensional bearbeiteten“ Beispieltabelle, Abb. 1.7, liefert das bemerkenswerte Ergebnis, dass die Anzahl der Sekundärsperrungen in dem betrachteten Streifen der aufgestockten Tabelle kleiner ist als in der durch Untertabellenabgleich „gesicherten“ zweidimensionalen Tabelle. Der Grund dafür liegt im Fall der aufgestockten Tabelle in der Gesamtsicht der Sicherungspartner, die zu einem vierdimensionalen Quader beitragen bzw. in der eingeschränkten Sicht bei Karreesicherung mit Untertabellenabgleich:

Für die linke unterste Untertabelle niedrigster Aggregation für sich alleine betrachtet (zweidimensionale Sicht) ist die Karreesicherung des Pivots (113; ACD) mit den Sekundärsperrungen in den Feldern (113; ACB), (110; ACB) und (110; ACD) mit dem besonders kleinen Eckwert 29 sicherlich günstiger als die bei vierdimensionaler Sicht in den entsprechenden Zeilen 113, 110 gewählten Felder der Spalten ACD und ACC. Doch erzwingt der Untertabellenabgleich aufgrund der Sperrung im Feld (110; ACB) eine Summensperrung in der Zeile 130, die im Inneren der obersten linken Untertabelle niedrigster Aggregation gegengesperrt werden muss. Außerdem fehlt die bei der Bearbeitung der vollständigen Tabelle erhaltene Summensperrung (130; ACC), die für die Sicherung des primär geheimen Feldes in der obersten Zeile schon ausgereicht hätte; so muss bei Untertabellenabgleich auch dieser Wert noch durch zusätzliche Sperrungen im Inneren der obersten linken Untertabelle gesichert werden.

Man sieht: Mit dem Aufstocken der Dimension, d.h. mit der Erhöhung des Schutzes der primär geheimen Werte ist nicht zwingend auch immer eine Erhöhung der Anzahl von Sekundärsperrungen verbunden. Es gibt vielmehr Tabellen, bei denen eine durch die Dimensionsaufstockung gewonnene Gesamtsicht zu weniger Sekundärsperrungen in der Veröffentlichungstabelle führen kann. Der erste Spaltenstreifen der Tabelle in Abb. 6.7 ist ein einfaches Beispiel dafür.

Das spaltenweise Vorgehen bei der Sicherung des zweiten Spaltenstreifens (ABC; ABB; ABA; ~~ABA~~; AB) führt zuerst zur Sicherung der Primärsperrung im Feld (123; ABB) mit dem Karree realer Tabellenwerte in den Feldern {(125; ABB), (125; AB), (123; ABB), (123; AB)} und den entsprechenden Projektionen in die Sternchensummentabellen. Durch diese Projektionen werden die beiden Sekundärsperrungen in den Zeilen 123 und 125 in der Randsummenspalte A verursacht. Ganz offensichtlich ist die Sperrung dieses Karrees bezüglich der Quaderwertesumme günstiger als das im zweidimensionalen Fall eingetragene Karree mit den Zwischensummen in der Zeile 120 (siehe Abb. 1.7), das durch die hierarchische Abarbeitung von Untertabellen nach absteigenden Aggregationsstufen erzwungen wurde.

Anders verhält es sich mit der zweiten Quadersicherung in dem mittleren Spaltenstreifen, die den primär geheimen Wert im Feld (112; ABA) betrifft. Hier ist die Zwischensummensperrung in der Zeile 110 durch die strukturellen Nullen vorbestimmt! Dadurch werden dann auch die beiden Sperrungen in die Randsumme, Zeile 100, hervorgerufen. Die Randsperrungen in der Spalte A und der Zeile 100 sind (bei fehlenden Randschranken) günstiger als die Projektion in andere Untertabellen mit realen Werten, weil die Projektion in die Sternchensummen mit realem Rand für jeden Quader immer noch insgesamt 10 negativ gewichtete Dummies in die Quadersumme einbringen. Außerdem wird mit dem Quader, der (112; ABA) schützt, auch die Primärsperrung (110; ABA) mitgesichert.

Die Randsummensperrungen ließen sich vermeiden, wenn man von der zu Anfang dieses Abschnitts angesprochenen Umsortierung der Dummy-Zeilen Gebrauch machen würde, wodurch die Primärsperrung in Zeile 112 in die dritte Zeile ihrer Untertabelle verlegt würde. Dann ließe sich ein vierdimensionaler Quader aufbauen mit denselben Sperrungen realer Werte wie im mittleren Spaltenstreifen der Abb. 1.7. Dieser Quader wäre hinsichtlich des Summenkriteriums wesentlich günstiger als jede der oben beschriebenen Sicherungen mit vierdimensionalen Quadern, weil dabei drei Primärsperrungen mit 8 Sternchensummen als Quaderwerte die Quadersumme beträchtlich verringerten. Mit dieser Optimierung durch Umordnung von Dummy-Zeilen ergeben sich bei Bearbeitung der als Ganzes zur vollständigen Tabelle aufgestockten Beispieltabelle drei Sperrungen weniger als beim Quaderverfahren mit Untertabellenabgleich, und das, ohne zusätzliche alternative Sperrungen in die Zwischen- bzw. Randsummenfelder in Kauf nehmen zu müssen.

Der dritte Spaltenstreifen in Abb. 6.7 (bzw. 6.8) hat beinah dieselbe Struktur gesperrter Tabellenfelder wie die Tabelle der Abb. 1.7; lediglich die „Gegensperrungen“ in der Zeile 112 zur Sicherung der beiden Primärsperrungen in der Zeile 113 in den Spalten AAD und AAB wären besser in die Zeile 111 verlegt worden, weil dadurch die Summe real zu sperrender Werte kleiner ausgefallen wäre. Dass hier trotzdem die etwas größeren Werte 745 und 67 als Sperrkandidaten ausgesucht wurden, lässt sich nur durch Betrachtung der in die Sternchensummen projizierten Tabellen verstehen und dadurch, dass gesperrte Sternchensummen stärker negativ in das Gesamtsummenkriterium des vierdimensionalen Sicherungsquaders eingehen als die noch „offenen“ Sternchensummen.

Das liegt an der dimensionsabhängigen Festlegung von geheimen Werten als Summanden in der Quadersumme und an der Wahl des Sternchensummenwertes und seines Gewichtes, mit dem er in die Quadersumme eingeht. In dieser vierdimensionalen Tabelle wird den geheimen Werten - 1, $1*(2^4-1)*100.000$ = - 1.650.000 in der Quadersumme zugeordnet (siehe Abschnitt 5.2.2), während der Klassenwert zu 60.000 gemäß ln60.000/ln61.000*99.999+1 = 99.850 (beachte obige Festlegungen der Steuerungsparameter) zu berechnen ist. Multipliziert mit dem Gewicht - 15 ergibt sich daraus der in die Quadersumme eingehende gewichtete Klassenwert zu 99.850*(-15) = - 1.497.750; er ist demnach um 152.250 größer und damit als noch durchzuführende Sperrung ungünstiger als eine bereits gesperrte Sternchensumme.

Vergleicht man damit die beiden zur Auswahl stehenden Quader, den in Abb. 6.8 eingetragenen mit dem, den man erhalten hätte, wenn man die beiden Werte in den Spalten AAD, AAB der Zeile 111 gesperrt hätte, so sieht man, dass die zuletzt genannten Sperrungen als Projektionen in die Sternchensummen mit 6 ungesperrten Sternchensummenwerten zur Quadersumme beigetragen hätten, während in Abb. 6.8 tatsächlich nur 3 ungesperrte Sternchensummen eingehen; die anderen drei Werte in den Feldern (112; ~~AAB~~), (~~112~~; ~~AAD~~) und (~~112~~; ~~AAB~~) sind bereits vorher gesperrt worden durch Projektion der Primärsperrung im Feld (112, ABA) und der Einzelangabe im Feld (133; ACD). Das ergibt 3*152.250 = 456750 Klassensummenpunkte weniger als bei der Sperrung gemäß Abb. 1.7. Diese Punktezahl wiegt den Unterschied zwischen den realen Summen, 745 + 67 - 148 - 81 = 583 bei weitem auf, sodass sich damit die in Abb. 6.7 bzw. 6.8 ausgeführten Sperrungen erklären.

Diese an sich nicht ganz befriedigende Lösung lässt sich verbessern, indem man die negativ gewichteten Klassenwerte der als Sperrpartner in Frage kommenden Dummywerte genauso groß wie die geheimen Werte der aufgestockten Tabelle in der Quadersumme wählt. Dazu genügt es, als Dummywert den größten Tabellenwert anzusetzen, dessen Klassenwert man dann bei Eintrag in die Quadersumme nur noch mit dem dimensionsabhängigen Faktor - 1, $1*(2^n - 1)$ (n bezeichnet die Dimension der aufgestockten Tabelle, vergleiche 5.2.2) zu gewichten hat; mit anderen Worten, jeder sperrbare Dummy erhält als gewichteten Klassenwert denselben Wert wie die geheimen Werte in der Quadersumme zuerkannt. Für die Beispieltabelle bedeutet das, dass die sperrbaren Dummies den gewichteten Klassenwert - 1.650.000 erhalten. Damit ergibt sich dann in dem dritten Spaltenstreifen der Abb. 6.7 dieselbe Sperrverteilung wie in der Tabelle der Abb. 1.7, weil sich nun die in die Sternchensummen projizierten Karrees mit vorherigen Sekundärsperrungen nicht mehr unterscheiden von denen, die noch keine oder weniger Sekundärsperrvermerke tragen.

Als Resümee dieses Abschnitts bleibt festzuhalten, dass das Quaderverfahren, das im Gegensatz zu einem allgemeinen Optimierungsverfahren zur Simultansicherung aller primär geheimen Werte eine Einzelpunktsicherung für jeden primär geheimen Wert durchführt, bei seiner Auswahl von Sekundärpositionen sehr wohl von der (sich während des Sperrvorgangs ändernden) Gesamtverteilung der gesperrten Werte geleitet wird. Dies ist bei der Bearbeitung der zur vollständigen Tabelle aufgestockten Beispieltabelle dadurch besonders deutlich geworden, dass die durch Dummy-Verlegungen optimierte Quadersicherung, ganz anders als erwartet, nicht zu mehr, sondern sogar zu weniger Sekundärsperrungen geführt hat als bei der mit Hilfe des Untertabellenabgleichs gesicherten zweidimensionalen Tabelle. Das konnte damit erklärt werden, dass eine Gesamtsicht von in Betracht kommenden Sicherungspartnern (über mehrere Untertabellen hinweg) die Anzahl von Übersperrungen reduzieren kann.

Umgekehrt bedeutet dies nun aber nicht unbedingt, dass eine Verkürzung der Tabelle aufgrund von sperrungsfreien Summen oder auch eine Vorwegnahme von Sperrungen auf niedrigstem Aggregationsniveau i. Allg. Übersperrungen begünstigen muss. Dem durch Tabellenverkürzung und Vorwegnahme von Sekundärsperrrungen zu erzwingenden beträchtlichen Gewinn an Rechenzeit steht u.U. lediglich eine gewisse Veränderung der Sperrverteilung gegenüber, ähnlich wie dies durch die Festlegung des Abarbeitungsschemas, z.B. zeilenweises, spaltenweises usw. Vorgehen, erfolgt. Es werden z.B. erst die auf unterstem Aggregationsniveau zu sichernden geheimen Werte durch n-dimensionale Quader geschützt (n bezeichnet die Dimension der aufgestockten Tabelle) und dann erst diejenigen, in denen auch höhere Aggregate vorkommen und die daher mehrere Untertabellen der Veröffentlichungstabelle überdecken, oder es erfolgt eine Abarbeitung nach absteigenden Verdichtungen, wie nachfolgend dargestellt.

6.2.2.3 Partielle Aufstockung zur Rechenzeitverkürzung

Zur genaueren Darstellung dieser wichtigen Methode zur Rechenzeitverkürzung durch Umstrukturierung des Sperrprozesses sei im folgenden zunächst der besonders leicht zu handhabende Spezialfall behandelt, bei dem die zu sichernden geheimen Werte auf unterstem Aggregationsniveau geschützt werden können. Dies betrifft genau diejenigen primär geheimen Werte, die bereits durch Quadersicherung in der ursprünglichen, noch nicht aufgestockten Tabelle innerhalb ihrer Untertabelle niedrigsten Aggregationsniveaus ohne Summensperrungen hinreichend zu sichern sind.

Die Gleichungssysteme dieser Sicherungsquader enthalten nur solche Unbekannten, die alle zur selben Untertabelle gehören und keine, die noch in Gleichungen anderer Untertabellen zu finden wären. Nach Aufstockung der Tabelle werden diese Sicherungsquader vollständig in die Sternchensummen projiziert, sodass keine zusätzlichen realen Sperrungen entstehen. Primär geheime Werte, die auf unterstem Aggregationsniveau in der ursprünglichen „unvollständigen“ Tabelle gesichert wurden, brauchen also in der vollständigen Tabelle nicht bearbeitet zu werden; es genügt, sie mitsamt ihren Sicherungspartnern als geheime Werte zu führen, die nicht mehr überprüft werden müssen, die aber bevorzugte Sicherungspartner für die Quaderauswahl in der vollständigen Tabelle darstellen.

Bei Verzicht auf Intervallschutz bietet sich zunächst an, die in einem Vorlauf in der unaufgestockten Tabelle auf unterstem Aggregationsniveau gesicherten primär geheimen Werte mit den zugehörigen Sicherungspartnern in der aufgestockten Tabelle als sperrbare Dummy-Werte zu behandeln und entsprechend zu markieren.

Durch die Möglichkeit der Vorabsicherung eines Teils von primär geheimen Werten auf unterstem Aggregationsniveau ist ein zweistufiges Vorgehen bei der Sicherung vollständiger Tabellen angezeigt: Die erste Stufe dient dazu, die auf unterstem Aggregationsniveau zu sichernden primär geheimen Werte aufzufinden und samt ihren Sicherungspartnern bezüglich der aufgestockten Tabelle als sperrbare Dummies zu markieren. Die zweite Stufe sichert dann die noch verbliebenen primär geheimen Werte in der aufgestockten vollständigen Tabelle, wobei die bereits auf erster Stufe gesicherten Werte und deren Partner als sperrbare Dummies besonders bevorzugte Sicherungspartner sind.

Dieses Vorgehen lässt sich nun zu einem mehrstufigen Prozess verfeinern, in dem alle aus einer Gesamttabelle zu extrahierenden Teiltabellen, die bezüglich jeder Gliederung eine Randsumme aufweisen und die immer auch die

untersten Aggregationsstufen einbeziehen, zu zwischensummenfreien Tabellen aufgestockt und dann nach absteigenden höchsten Aggregationsstufen mit dem Quaderverfahren partiell gesichert werden: Partielle Sicherung bedeutet, dass in jeder Teiltabelle immer nur für diejenigen Primärsperrungen ein Quader mit der Dimension der betreffenden aufgestockten Teiltabelle aufzusuchen ist, die in dieser Teiltabelle die höchsten Aggregationsstufen aufweisen und deren Quader im Inneren der betreffenden aufgestockten Teiltabelle liegen.

Das „Innere einer Teiltabelle“ bezeichnet diejenigen Tabellenfelder, die nicht den die Teiltabelle abtrennenden Randsummen angehören. Da jede solche Teiltabelle durch genau eine Untertabelle höchster Aggregationsstufen gekennzeichnet ist – sie ist durch eine Randsumme in jeder Gliederung abgeschlossen -, kann die Abarbeitung bzw. die Organisation der Teiltabellen genauso erfolgen, wie die der Untertabellen, durch Abarbeitung nach abnehmenden Aggregationsstufen und aller Positionsindizes innerhalb eines Satzes von Aggregationsstufen.

Bei der Geheimhaltung mit Intervallschutz tritt hier eine Schwierigkeit auf, die sich auf die Sonderbehandlung von sperrbaren Dummies gründet: Sperrbare Dummy-Werte werden im Allgemeinen durch einheitliche große positive Tabellenwerte mit ebenfalls einheitlichen negativen Gewichten ersetzt; sie haben keine Wertinformation, die bei der Berechnung von Quaderspannweiten verwendbar wäre. Im Summenkriterium werden sie mit geheimen Werten gleichgesetzt (vergleiche 5.2.2).

Die Bevorzugung geheimer Quaderwerte vorangegangener Sicherungen kann aber auch mit Erhalt der Wertinformation geschehen, indem der dem Wert entsprechende offene Klassenwert im Realteil und der Kehrwert dieses Klassenwertes multipliziert mit dem Wert, den ein geheimer Wert in der Quadersumme hat, als Gewicht im Imaginärteil der komplexen Wertvariablen eingetragen wird (siehe dazu 5.3, Justierung durch externe Gewichtung). Dann ist das in die Quadersumme als Summand einzufügende Produkt aus Klassenwert und Gewicht (Realteil * Imaginärteil) der einheitliche Wert aller geheimen Werte, während sich die Wertinformation aus dem im Realteil abgespeicherten Klassenwert ergibt.

Diese Spezialbehandlung muss aber nur auf die besonders gekennzeichneten Sperrvermerke, bei den als Dummies gesetzten real vorhandenen Tabellenwerten, angewendet werden, die anderen „gewöhnlichen“ sperrbaren Dummies und Sternchensummen werden weiterhin mit einheitlichen Gewichten und Werten belegt. Ferner muss man beachten, dass die dabei vorzunehmende Gewichtung nicht von vorneherein in den Datenbestand eingetragen werden kann, sondern fortlaufend abgespeichert werden muss. Sollen darüber hinaus auch noch Schätzintervalle berücksichtigt werden, so benötigt man eine EDV-Programmversion mit doppeltgenauem komplexen Wertefeld (vgl. Übersicht zur Struktur des Wertefeldes der Gesamttabelle im Hauptspeicher am Ende von 3.2.3.2). Außerdem muss man bei jeder Quadersummenbildung das negative Vorzeichen der bevorzugten Sicherungspartner noch gesondert berücksichtigen.

Schließlich kann man auf der Suche nach weiteren rechenzeit- und hauptspeicherplatzsparenden Aufstockungsverfahren – wie am Ende von 6.2.2.1 bereits angedeutet – ganz auf die Erstellung einer aufgestockten Tabelle im Hauptspeicher oder auf anderen Datenträgern verzichten. Statt dessen wird die vollständige Tabelle als Modellvorstellung genutzt, um für jeden primär geheimen Wert einen Sicherungsquader mit hinreichendem Intervallschutz auszuwählen, ohne dabei immer den gesamten hochdimensionalen Raum der vollständigen Tabelle abtasten zu müssen. Als angenehmer Nebeneffekt kann auf die oben beschriebene besondere Gewichtung von geheimen Wer-

ten aus vorangegangenen Sicherungen verzichtet werden – wie aus den folgenden Ausführungen zu entnehmen sein wird - , sodass das Gewichtsfeld wieder zur freien Verfügung steht.

Mit der Hilfskonstruktion „fiktive vollständige Tabelle“ lässt sich für jeden primär geheimen Wert – ganz individuell – eine kleinste Teiltabelle so finden, dass zumindest ein Sicherungsquader im Inneren dieser Teiltabelle liegt und jede weitere Verkleinerung der Teiltabelle immer zu Sperrungen in den Überlappungsbereich mit dem Rest der Statistiktabelle führt. Der für den betreffenden geheimen Wert nach alternativen Sicherungsquadern abzusuchende Raum beschränkt sich damit auf diese minimale Teiltabelle, andere evtl. existierende „Minimaltabellen“ bleiben dabei außer Acht. Dieses Verfahren wird im Folgenden weiter ausgeführt, wobei auch technische Details anzusprechen sind.

7. Quaderverfahren in fiktiver vollständiger Tabelle

7.1 Aufbau einer fiktiven vollständigen Tabelle

Die fiktive Tabelle wird durch ihre Elementarindizes simuliert. Den Bezug zu der gegebenen Statistiktabelle stellen Indexreferenz-Tabellen her, von denen es für jede reale Gliederung genau eine gibt. Alle Indexreferenz-Tabellen fasst eine Indexreferenz-Datei zusammen. Für den Aufbau und die Nutzung der Indexreferenz-Datei ist hilfreich, dass jede einem Gliederungskriterium der realen Tabelle zugeordnete Indexreferenz-Tabelle unabhängig von allen anderen erzeugt werden kann. Eine einmal aufgebaute Indexreferenz-Tabelle – z.B. die für die vielen Statistiktabellen gemeinsame regionale Gliederung – lässt sich in nachfolgenden Anwendungen des Quaderverfahrens immer wieder einsetzen, weil ihre Indizes und deren Struktur allein durch die jeweilige gegebene reale Gliederung, nicht aber durch andere Gliederungen der Statistiktabelle bestimmt ist. Im Folgenden braucht man daher nur eine einzelne Indexreferenz-Tabelle zu betrachten.

7.1.1 Aufbau einer Indexreferenz-Tabelle zu vorgegebenem Nutzerindex

Gegeben sei eine mehrstufige hierarchische Gliederung, die durch ihren Nutzerindex repräsentiert wird. Dieser mehrstufige Index sei bereits durch die zugehörigen Aggregations- und Positionsindizes ergänzt (vgl. Abschnitt 1.2.3). Zur Illustration wird hier ein vierstufiger Nutzerindex als Beispiel mitgeführt.

Abb. 7.1: Nutzerindex mit vier Aggregationsstufen

mehrstufiger Index	1	2 3 4	5	6 7	8	9	10 11 12	13 14	15	16
Aggregations-Index	2	1 1 1	2	1 1	2	3	1 1 1	2 2	3	4
Positions-Index	1	1 1 1	1	2 2	1	1	3 3 3	2 2	1	1

Als reales Beispiel kann auch die regionale Gliederung in NRW dienen, mit den Gemeinden auf der ersten, untersten Aggregationsstufe, den Kreisen und kreisfreien Städten auf der zweiten, den Regierungsbezirken auf der dritten und dem Land auf der vierten, höchsten Aggregationsstufe.

Zu jeder Aggregationsstufe a mit a = 1, 2, 3, ..., A, A = höchste Aggregationsstufe (Randsumme), gibt es eine (nicht leere) Gesamtheit von Positionsindizes. Alle Nutzerindizes zur Aggregationsstufe a, die alle den selben

Positionsindex haben, werden zu einem Segment (dieser Gliederung und) dieser Aggregationsstufe zusammengefasst.

Im Nutzerindex-Beispiel bezeichnen die Indexmengen {2, 3, 4}, {6, 7}, {10, 11, 12} die drei Segmente zur ersten Aggregationsstufe (dieser Gliederung), die Gesamtheiten {1, 5, 8}, {13, 14} die Segmente der zweiten Aggregationsstufe, {9, 15} das einzige Segment der dritten Stufe und {16} das einzige Segment der höchsten, der vierten Aggregationsstufe.

Die Anzahl der Elemente eines Segments wird als dessen Segmentlänge bezeichnet. Da es nur eine Randsumme (mit der Aggregationsstufe A) gibt, existiert für die beiden höchsten Aggregationsstufen a = A, A–1 immer jeweils nur ein Segment bzw. jeweils nur eine Ausprägung für den zugehörigen Positionsindex. Die Positionsindizes zu den Aggregationsstufen A und A–1 haben daher als Indizes zur Unterscheidung von Segmenten keinen Sinn, sie sind demnach für die Aufstockung der Tabellendimension irrelevant.

Für die Quaderauswahl sind ferner die Indexabschnitte von Bedeutung: Ein Indexabschnitt zur Aggregationsstufe a_s bezeichnet die Gesamtheit aller Nutzerindizes, die zu einem Segment der Aggregationsstufe a_s gehören und alle Indizes zu Aggregationsstufen $a < a_s$, die zu den Indizes des Segments zu a_s beitragen. Das sind die Nutzerindizes des Segments zu a_s und alle Nutzerindizes mit $a < a_s$, die dem jeweiligen Nutzerindex zu a_s in der Sortierung unmittelbar vorausgehen, zurück bis zu dem ersten Index zur Aggregationsstufe a_u, für die $a_u > a_s$ gilt (Sortierungskonvention gemäß Einführung).

Wie oben bereits angedeutet, liefern alle zu den Aggregationsstufen $a \leq A\text{-}2$ gehörigen Positionsindizes zusammen mit den Nutzerindizes der längsten Segmente unterster Aggregation bereits die A–1 Elementarindizes, nach denen die „dimensionsaufgestockte“ vollständige Tabelle bezüglich des betrachteten Gliederungskriterium gegliedert ist. Was fehlt, sind die für die Auffüllung der Segmente jeder Aggregationsstufe zur vollen maximalen Segmentlänge einzufügenden Dummies sowie die Sternchensummen. Um auch diese zu berücksichtigen, wird im Folgenden ein mechanisiertes Vorgehen skizziert, mit dem, beginnend mit der untersten Aggregationsstufe a = 1, ein Elementarindex nach dem anderen – und das unabhängig von den anderen Elementarindizes – aufgebaut wird, ohne dabei die Dummy-Indizes eintragen zu müssen.

7.1.1.1 Aufbau eines Elementarindexes zur a–ten Aggregationsstufe

Jeder Elementarindex erstreckt sich immer nur über zwei Aggregationsstufen, über die unterste und über die Randsummen-Aggregation. Die unterste und damit auch die erste Aggregationsstufe zum a–ten Elementarindex ist die a–te Aggregationsstufe des umzuindizierenden Nutzerindex, a = 1, 2, ..., A–1. Der a-te Elementarindex wird nach folgendem Schema schrittweise aufgebaut:

α) Bestimmung der maximalen Segmentlänge der a-ten Aggregationsstufe

Zur vorgegebenen Aggregationsstufe a wird aus der Nutzerindextabelle ein Segment nach dem anderen aufgesucht und seine Segmentlänge bestimmt. Dabei werden alle Positionsindizes zur Aggregationsstufe a durchlaufen. Aus all diesen Segmentlängen $L_{i,a}$, i = 1, 2,..., I_a, I_a = höchster Positionsindex zur Aggregationsstufe a, wird die größte, L_a, ausgewählt.

β) Durchnummerieren der Nutzerindizes innerhalb der Segmente einschließlich ihrer Randsummen zur Aggregationsstufe a von 1 bis L_a+1 und Eintragung in die Indexreferenz-Tabelle

Die Nummernfolge $\{E_a\} = \{1, 2 ..., L_a\}$ bezeichnet den Elementarindex unterster Aggregation zur a-ten Aggregationsstufe des Nutzerindexes. Dabei hat man sich diejenigen Segmente mit weniger als L_a Nutzerindizes durch Dummy-Werte ergänzt vorzustellen. (Diese Art Dummies dürfen nicht als Quaderwerte dienen, es sei denn, alle Werte des Segments sind Dummies.) Der a-te Elementarindex mit Randsummenindex ist dann

$$\{E_a\} = \{1, 2, ..., L_a, L_a+1\}.$$

Beim Durchnummerieren der Nutzerindizes in allen Segmenten einschließlich ihrer Randsummen, d.h. von 1 bis L_a+1, wird der dem jeweiligen Nutzerindex (falls vorhanden) entsprechende Elementarindex direkt in die Indexreferenz-Tabelle eingetragen. Dabei erhalten alle Nutzerindizes mit Aggregationsstufe > a den Elementarindex L_a+1 zuerkannt, weil sie eine höhere Aggregationsstufe als das „Innere“ der gerade zu bearbeitenden Segmente haben. Alle Nutzerindizes zu Aggregationsstufen < a erhalten den selben Elementarindex wie der Nutzerindex zu a, zu dem diese beitragen.

7.1.1.2 Zur Aufstellung der Indexreferenz-Tabelle zu einem gegebenen Nutzerindex

γ) Allgemeines Vorgehen

Beginnend mit der untersten Aggregationsstufe a = 1 wird in einer Schleife zum Index a = 1, 2, ..., A-1 ein Elementarindex $\{E_a\}$ nach dem anderen gemäß Punkt 7.1.1.1 aufgebaut. Dabei werden gleichzeitig die Elementarindizes innerhalb jedes Segments einschließlich ihrer Randsummen als Nummern E_a = 1, 2, ..., L_a, L_a+1 der Segmentelemente den zugehörigen Nutzerindizes zugeordnet – sofern diese Nutzindizes vorhanden sind. Dieses Vorgehen ist zulässig, weil bei jeder Aufstellung eines neuen Elementarindexes immer nur von den zugehörigen Nutzerindizes der betroffenen Aggregationsstufe Gebrauch gemacht wird, nicht aber von den vorangegangenen Elementarindizes.

Anmerkung:

Die Abhängigkeit der Elementarindex-Zuordnung von den Zuordnungen der in den vorangegangenen Schritten bestimmten Elementarindizes tritt erst dann in Erscheinung, wenn die Dummy- und Sternchensummen-Indizes real mitgeführt werden sollen. Diese Abhängigkeit ist insbesondere dann von Bedeutung, wenn man eine aufgestockte Tabelle auf Datenträger erstellen will – wie bisher bei Anwendungen von GHQUAR.4 praktiziert (vgl. 6.2.2.2 und Anhang A.4).

Aus der (real) aufgestockten Beispieltabelle der Abbildung 6.8 (Pkt. 6.2.2.2) entnimmt man die Elementarindizes zum Nutzerindex der Spalten:

Nutzerindex der Spalten in der Beispieltabelle von Abb. 6.8 (mit Dummy-Eintragungen)	ACD	ACC	ACB	ACA	AC	ABC	ABB	ABA	D	AB	AAD	AAC	AAB	AAA	AA	SD	SD	SD	SD	A
1. Elementarindex	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
2. Elementarindex	1	1	1	1	1	2	2	2	2	2	3	3	3	3	3	4	4	4	4	4

Wenn darin die Dummies D und die Sternchensummen SD weggelassen werden, erhält man die hier beschriebene verkürzte Indexreferenztabelle.

Um den Aufstockungsmechanismus nicht unnötig zu verkomplizieren, wurde auf das Mitführen von Sternchensummen und Dummies in nachfolgend abgehandeltem Beispiel verzichtet.

7.1.1.3 Aufstockung eines vierstufigen hierarchischen Beispiel-Indexes

Als Beispiel-Index dient der Nutzerindex der Abbildung 7.1 mit den dort angegebenen Aggregations- und Positions-Indizes. Nach Punkt γ wird zunächst der Aggregationsschleifenindex a auf a = 1 gesetzt und mit den a = 1 entsprechenden Segmenten {2, 3, 4} zur Position 1, {6, 7} zur Position 2 und {10, 11, 12} zur Position 3 gemäß α die maximale Segmentlänge zu $L_1 = 3$ ermittelt. Der erste Elementarindex ergibt sich dann gemäß β zu

$$\{E_1\} = \{1, 2, 3, 4\}.$$

Um die Elemente dieses Elementarindex den entsprechenden Nutzerindex-Elementen zuzuordnen und dann gemäß β in die Indexreferenz-Tabelle einzutragen, werden zunächst die Indizes der Segmente zu a = 1 einschließlich ihrer Randsummen von 1 bis 4 durchnummeriert. Diese Nummern wurden in nachstehender Abbildung 7.2 als 1. Elementarindex eingetragen und durch Fettdruck besonders herausgehoben. Übrig bleiben nun noch Indizes zu Aggregationsstufen > a = 1; sie können aus der „Sicht“ des ersten Elementarindex nur als Randsummen in Erscheinung treten und erhalten daher den Elementarindex $L_1 + 1 = 4$ zuerkannt.

Abb. 7.2: Indexreferenztabelle mit erstem Elementarindex

mehrstufiger Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Aggregations-Index	2	1	1	1	2	1	1	2	3	1	1	1	2	2	3	4
Positions-Index	1	1	1	1	1	2	2	1	1	3	3	3	2	2	1	1
1. Elementarindex	4	**1**	**2**	**3**	**4**	**1**	**2**	**4**	4	**1**	**2**	**3**	**4**	4	4	4

Der Aggregationsschleifenindex a wird gemäß γ um 1 erhöht und mit dem höchsten Aggregationsindex A = 4 verglichen; da a = 2 immer noch kleiner als A ist, wird nach α mit der Aufsuche der Segmente zu a = 2 fortgefah-

ren: {1, 5,8}; {13, 14}. Die größte Segmentlänge zu a = 2 beträgt demnach $L_2 = 3$. Der zweite Elementarindex $\{E_2\}$lautet entsprechend β

$$\{E_2\} = \{1, 2, 3, 4\}$$

Dieser Index wird – wiederum nach β – als zweiter Elementarindex in die Indexreferenztabelle (Abb. 7.3) eingebracht, wobei wieder alle Nutzerindizes innerhalb der Segmente zu a = 2 einschließlich ihrer Randsummen von 1 bis 4 durchnummeriert und diese Nummern in Fettdruck in die Indexreferenzdatei eingetragen werden. Außerdem erhalten alle Nutzerindizes mit Aggregationsstufen größer als a = 2 den Elementarindex-Eintrag $L_2 + 1 = 4$.

Bei diesem Schleifendurchlauf kommen noch die mit Elementarindizes zu versehenden Nutzerindizes zu Aggregationsstufen < a = 2 neu hinzu; sie erhalten denjenigen Elementarindex als Eintrag, zu dem diese Aggregate beitragen, also den jeweils nachgeordneten (hier fett gedruckten) Elementarindex der Aggregationsstufe a = 2.

Abb. 7.3: Indexreferenztabelle mit allen drei Elementarindizes

mehrstufiger Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Aggregations-Index	2	1	1	1	2	1	1	2	3	1	1	1	2	2	3	4
Positions-Index	1	1	1	1	1	2	2	1	1	3	3	3	2	2	1	1
1. Elementarindex	4	**1**	**2**	**3**	**4**	**1**	**2**	**4**	4	**1**	**2**	**3**	**4**	4	4	4
2. Elementarindex	**1**	2	2	2	**2**	3	3	**3**	**4**	1	1	1	**1**	**2**	**4**	4
3. Elementarindex	1	1	1	1	1	1	1	1	**1**	2	2	2	2	2	**2**	**3**

Zum letzten Mal wird der Aggregationsschleifenindex a um 1 erhöht (gem. γ), da dieser danach mit a = 3 gerade noch kleiner als A = 4 ist, und es wird (nach β) das einzige Segment zu a = 3 aufgesucht: {9, 15}; $L_3 = 2$. Mit β ergibt sich daraus der dritte Elementarindex $\{E_3\}$ zu

$$\{E_3\} = \{1, 2, 3\}.$$

Nach Durchnummerieren der Indizes innerhalb des Segmentes zu a = 3 einschließlich seiner Randsumme erhält man nach β die fett gedruckten Elementarindex-Ausprägungen in der letzten Zeile der Indexreferenztabelle (Abb. 7.3). Abschließend werden in diese Zeile noch die Elementarindizes eingetragen, die mit den jeweils nachgeordneten, fettgedruckten Elementarindizes zur Aggregationsstufe a = 3 übereinstimmen. Der Aufbau der Indexreferenztabelle ist damit abgeschlossen.

7.1.2 Auswahl von Sicherungsquadern mit der Indexreferenz-Datei

Beim Aufbau von Sicherungsquadern in der von den Elementarindizes aufgespannten vollständigen Tabelle muss man zwei wesentliche Aspekte berücksichtigen, die Unterscheidung und Bewertung von erlaubten und verbotenen Dummies und die Beschränkung der Auswahl von Quaderelementen auf durch Indexabschnitte aus der Gesamttabelle abgegrenzte Teiltabellen. Beide Aspekte wurden oben bereits eingehend diskutiert; hier wird nur noch die technische Umsetzung erläutert.

7.1.2.1 Unterscheidung und Bewertung von Dummies

δ) Verbotene Dummies

Ein Dummy ist verboten, d.h. als Element eines Sicherungsquaders unzulässig, wenn er bezüglich einer der Nutzerindextabellen verboten ist. Ein Dummy ist bezüglich einer Nutzerindextabelle verboten, wenn sein Nutzerindex mit anderen vorhandenen Nutzerindizes (der selben Aggregationsstufe) zum selben Segment beiträgt. Alle anderen Dummies sind erlaubt, d.h. als Elemente eines Sicherungsquaders zulässig. Dieses „Dummy-Verbot“ geht über das in Abschnitt 6.2.2.1 etwas hinaus. Es verbietet auch Dummies, die – wie in Abb. 6.8 im mittleren Spaltenstreifen – im Schnittbereich zweier oder mehrerer eingefügter Gliederungen liegen. Diese Dummies sind als Sperrpartner zwar prinzipiell erlaubt, sie können aber immer nur zu Quadern gehören, die strukturelle Nullen oder verbotene Dummies enthalten, was keinen Einfluss auf das Sperrmuster einer Tabelle haben kann.

Um die Entscheidung, ob es sich um einen verbotenen Dummy handelt oder nicht, bereits allein in der (fiktiven) vollständigen Tabelle treffen zu können, muss man obige Definitionen auf die Elementarindizes übertragen:

Ein Dummy ist verboten, wenn einer seiner Elementarindizes mit anderen in der Indexreferenz-Datei aufgeführten (also real existierenden) Elementarindizes zum selben Randsummen-Elementarindex gehört, alle anderen Dummies sind erlaubt.

Folgende Dummies einer Statistiktabelle, deren einer Nutzerindex mit dem der Abbildung 7.1 identisch ist, sind auf jeden Fall verboten, wie immer sich auch die Elementarindizes zu den anderen Nutzerindizes verhalten (vgl. Abb. 7.3):

$(E_1, E_2, E_3) = (3, 3, 1)$ 2. Segment zur ersten Aggregationsstufe wird ergänzt

$(E_1, E_2, E_3) = (4, 3, 2)$ 2. Segment zur zweiten Aggregationsstufe wird ergänzt

Folgende Dummies der Beispieltabelle (Abb. 7.3) sind erlaubt, wenn sie nicht durch die anderen Indexreferenztabellen verboten sind:

$(E_1, E_2, E_3) = (2, 1, 1)$ „Hinterland“ erster Aggregationsstufe wird ergänzt

$(E_1, E_2, E_3) = (1, 2, 2)$ „Hinterland“ erster Aggregationsstufe wird ergänzt

Beim Durchlaufen einer jeden Elementarindexschleife muss man sich merken, ob der vorhergehende Elementarindexwert einen realen Tabellenwert indiziert oder nicht; wenn nicht, ist der Dummy erlaubt, es sei denn die anderen Elementarindizes verbieten ihn, wenn ja, ist der Dummy verboten und es muss sofort ans Schleifenende dieses Elementarindex gesprungen werden, weil keiner der noch folgenden Elementarindexwerte zu einem erlaubten Dummy führt.

Bei dieser Vorgehensweise ist die fiktive vollständige Tabelle so sortiert, dass in jeder Elementarindex-Dimension zuerst die realen Tabellenfeldern entsprechenden Indizes stehen und erst dann die der Dummies.

η) Bewertung von erlaubten Dummies

Bei der Auswahl eines Sicherungsquaders müssen auch erlaubte Dummies noch bewertet werden und zwar nur in Bezug auf das Summenkriterium, für die Berechnung des Quaderschutzintervalls sind Dummies ohne Belang. Beim Eintrag der erlaubten Dummies in die Quadersumme könnte man diese - mit den gleichen Argumenten wie in Abschnitt 6.2.2.2 - mit den selben negativen Zahlen belegen wie die geheimen Werte. Die Erfahrung mit Realdaten hat aber gezeigt, dass mit so einer Bewertung zwar eine besonders kleine Zahl von Sekundärsperrungen zu erreichen ist, dass dies aber auf Kosten einer großen Zahl von Summensperrungen geht. Die Bevorzugung von Randsummen als Sperrkandidaten erklärt sich dadurch, dass Sternchensummen als Quaderwerte andere Quaderwerte gleichen Aggregationsniveaus nach sich ziehen und das sind eben häufig auch offene reale Randsummen.

In Abschnitt 6.2.2.2 war eine stark negative Bewertung der Sternchensummen vorgenommen worden, um damit die im Inneren einer Teiltabelle gelegenen niedrigdimensionalen Quader vollständig in die Sternchensummen zu projizieren. Bei der hier nur fiktiv vorzunehmenden Aufstockung, bei der ausschließlich diejenigen Randsummen in den Quaderauswahlprozess einbezogen werden, ohne die eine Sicherung des gerade betrachteten Pivots nicht möglich wäre, stellt sich dieses Problem gar nicht:

Während in der real aufgestockten 4-dimensionalen Beispieltabelle der primär geheime Wert im Tabellenfeld (133; ACD) durch den vierdimensionalen, $2^4 = 16$ Tabellenwerte umfassenden Quader {(134; ACD), (134; ACC), (133; ACD), (133; ACC), (~~113;~~ ACD), (~~113~~; ACC), (~~112~~; ACD), (~~112~~; ACC), (134; ~~AAD~~), (134; ~~AAC~~), (133; ~~AAD~~), (133; ~~AAC~~), (~~113~~; ~~AAD~~), (~~113~~; ~~AAC~~), (~~112~~; ~~AAD~~), (~~112~~; ~~AAC~~)} gesichert werden muss, begnügt man sich bei fiktiver Aufstockung mit den ersten vier realen Werten. Um Sperrungen in andere reale Untertabellen zu unterbinden, muss eine Projektion in die Sternchensummen gar nicht erst erzwungen werden, sie findet einfach nicht statt, ist von vorneherein ausgeschlossen.

Damit wird die Bewertung der Sternchensummen und der „Hinterland-Dummies“ für die Vermeidung von Summensperrungen bedeutsam: sie kann im Falle der fiktiven Aufstockung positiv und vergleichbar mit den gleichrangigen realen Randsummen gewählt werden, so dass ein solche Randsummen meidender Quader eine kleinere Quadersumme hat als einer mit realen Randsummen und Sternchensummen des selben Aggregationsniveaus. Diese Bewertung ist abhängig vom Aggregationsniveau des betreffenden Summenwertes und damit von der Anzahl der Elementarindizes, welche für die Einbeziehung des jeweiligen Summenfeldes in den Sicherungsquader erforder-

lich sind. Um beispielsweise das Randsummen-Dummyfeld (~~113~~; AB) der Beispieltabelle, Abb. 6.8, zu bewerten, sind in der Spaltendimension der ursprünglichen Tabelle zwei Elementarindizes, in der Zeilendimension ein Elementarindex zu berücksichtigen.

7.1.2.2 Auswahlbereiche für Sicherungsquader

λ) Allgemeine Beschreibung von Auswahlbereichen

Die Vervollständigung einer Statistiktabelle durch Aufstocken der Tabellendimension erweitert den Auswahlbereich von Sicherungsquadern zum Schutze primär geheimer Werte ganz enorm. Damit verbunden ist auch eine enorme Rechenzeitzunahme (siehe insbesondere Anhang A.4). Um – zum Zwecke der Rechenzeitreduktion – nicht mehr den gesamten Raum der vollständigen Tabelle abtasten zu müssen und dabei mit dem Quaderverfahren trotzdem einen hinreichenden Intervallschutz garantieren zu können, genügt es, bei der Sicherung eines geheimen Wertes nur solche Quader zu untersuchen, die gewisse Zwischensummen in der realen Tabelle meiden, deren Auswahlbereich also durch eben diese Zwischensummen eingegrenzt wird.

Es wird also für jeden primär geheimen Wert – ganz individuell – ein Auswahlbereich durch Zwischensummen niedrigster Aggregation abgegrenzt und darin werden dann alle Quader zum Schutze des einen geheimen Wertes untersucht. Die für die Abgrenzung des Auswahlbereiches in Frage kommenden „Zwischensummen niedrigster Aggregation“ sind dadurch ausgezeichnet, dass sie keine Quaderelemente eines der Sicherungsquader des betrachteten zu sichernden geheimen Feldes enthalten und von allen solchen Zwischensummen die niedrigste Aggregationsstufe aufweisen. Erst, wenn dieser vor der Sicherung des primär geheimen Wertes abzugrenzende Auswahlbereich für die Quaderauswahl nicht mehr ausreicht – zu viele Leerstellen, zu viele verbotene Dummies, zu kleine Schutzintervalle usw. -, wird der Auswahlbereich bis hin zu den entsprechenden Zwischensummen der nächst höheren Aggregation ausgeweitet.

Auswahlbereiche umfassen immer auch alle unteren Aggregationsstufen bis hinauf zu den Aggregationsstufen der begrenzenden Zwischensummen, sie sind daher Teiltabellen, die durch die zugehörigen Indexabschnitte festgelegt sind. Wie jede Untertabelle wird auch jede Teiltabelle durch ihre Aggregations- und Positionsindizes eindeutig bestimmt. Anders als die Untertabelle umfasst die Teiltabelle aber auch alle vorausgehenden, zu den höheren Aggregaten beitragenden Tabellenwerte bis zur untersten Aggregation.

Um die Teiltabellen bezüglich der Elementarindizes festzulegen, kann man daher die Hierarchie der Elementarindizes bezüglich der Aggregationsniveaus, aus denen sie hervorgehen, nutzen. Es genügt hier, statt der Teiltabellen nur die Indexabschnitte einer Gliederung zu untersuchen: Der in der gegebenen Gliederung durch den Aggregationsindex a und den Positionsindex i eindeutig festgelegte Indexabschnitt wird im Raum der Elementarindizes durch alle (konstanten) Elementarindizes zu Aggregationsstufen > a eindeutig festgelegt, wobei die Elementarindex-Stufen > a der Aggregationsstufe a und die Ausprägungen der höheren konstanten Elementarindizes der Position i in der gegebenen Teiltabelle entsprechen (vgl. Abb. 7.3). Die Aggregations- und Positionsindizes werden

demnach nur für die Aufstellung der Indexreferenztabellen benötigt, danach sind sie entbehrlich und brauchen daher nicht mehr mitgeführt zu werden.

Um für einen gegebenen zu sichernden geheimen Wert unter den vielen Teiltabellen eine geeignete als Auswahlbereich zu finden, sucht man zunächst nach einem niedrigdimensionalen Quader, indem in allen Nutzergliederungen nur die untersten Elementarindizes variiert werden, alle höheren bleiben fest. Gelingt das nicht, weil ein Quaderelement in eine begrenzende Zwischensumme fällt, d.h. den höchsten Indexwert L_a+1 des betrachteten Elementarindex annimmt, wird bezüglich der betroffenen Nutzergliederung der nächst höhere Elementarindex hinzugenommen und dann alle Elementarindizes niedrigerer Aggregation dieser erweiterten Teiltabelle variiert usw.

Hat man schließlich einen Satz von eine Teiltabelle fixierenden Elementarindizes (der höheren Aggregationsstufen, die für alle Elementargliederungen dieser Teiltabelle fest bleiben) gefunden, für den ein Sicherungsquader für das zu schützende geheime Tabellenfeld existiert, so ist die dadurch abgegrenzte Teiltabelle bereits ein geeigneter Auswahlbereich. Diese Teiltabelle wird nach weiteren Quadern durchsucht und daraus der nach den bekannten Auswahlkriterien günstigste als Sicherungsquader ausgewählt.

- Beim Aufbau von Sicherungsquadern in den niedrigdimensionalen Teiltabellen erfolgt die Auswahl der diese Quader fixierenden Diametralelemente nur in Bezug auf die wenigen Elementarindizes niedrigster Aggregation, die den Auswahlbereich aufspannen. Alle Elementarindizes, die den höheren Aggregationsstufen angehören, die in der betrachteten Teiltabelle nicht mehr variieren, sind davon ausgenommen (und stimmen infolgedessen mit denen des zu schützenden Wertes überein). Ein Diametralelement in einer niedrigdimensionalen Teiltabelle ist niemals auch Diametralelement zum zu schützenden Wert im von allen Elementarindizes aufgespannten Tabellenraum. –

μ) Auswahlbereiche im Nutzerindexbeispiel

μα) Das Pivot-Element sei im Beispielindex der Abb. 7.3 durch den Nutzerindex 6 gekennzeichnet

Dem entspricht das Elementarindex-Tripel (1, 3, 1). Die zu diesem zu sichernden Wert gehörige Teiltabelle niedrigster Aggregation hat bezüglich des Beispiel-Nutzerindex den durch den konstanten 2. und 3. Elementarindex $E_2 = 3$, $E_3 = 1$ gekennzeichneten Indexabschnitt mit den Nutzerindizes 6, 7, 8 und den Elementarindizes (1, 3, 1), (2, 3, 1), (3, 3, 1) = verbotener Dummy, (4, 3, 1) = Summe. Ist der Tabellenwert mit Elementarindextripel (2, 3, 1) mit allen Quaderkriterien verträglich, so hat man bereits bezüglich des Beispielindex den gesuchten Auswahlbereich gefunden; er ist in diesem Beispielindex durch die zwei Elementarindizes (E_2; E_3) = (3, 1) gekennzeichnet. Die Auswahlschleife betrifft nur die Variation des ersten Elementarindex E_1 = 1, 2, 3 (ohne die Summe).

Beim Aufsuchen von zu sichernden primär geheimen Werten bedient man sich zweckmäßig der Nutzerindizes und geht erst nach dem Auffinden des Pivots in die aufgestockte Tabelle, wie es in diesen Beispielen angedeutet wurde. Um dabei nicht jedes Mal am Anfang aufzusetzen, merke man sich die Nutzerindizes des jeweils letzten zu schützenden primär geheimen Wertes.

Genügt der Tabellenwert (2, 3, 1) nicht den Quaderkriterien, so bleibt in dieser untersten Schleife nur noch der Dummy (3, 3, 1) und der ist gemäß δ verboten. Daraus folgt, dass der nächst höhere Elementarindex mit in den Auswahlbereich einbezogen werden muss: Der Auswahlbereich ist bezüglich des Beispielnutzerindex nur noch durch den konstanten dritten Elementarindex $E_3 = 1$ fixiert und überdeckt somit alle Nutzerindizes von 1 bis 8 mit 9 als Randsumme (siehe Abb. 7.3).

μβ) Das Pivot-Element sei durch den Nutzerindex 14 gekennzeichnet

Dem entspricht das Elementarindex-Tripel (4, 2, 2). Der erste Elementarindex des Pivots liegt wegen $E_1 = L_2+1 = 4$ bereits im Rand des untersten Aggregationsniveaus, so dass auch der zweite Elementarindex variiert werden muss: Der entsprechende Auswahlbereich (bezüglich dieser Gliederung) ist demnach nur durch den konstanten dritten Elementarindex $E_3 = 2$ fixiert, die beiden anderen, E_1, E_2 spannen den Auswahlbereich auf; sie sind beide zu variieren, und zwar E_1 von 1 bis $L_1+1 = 4$ und E_2 von 1 bis $L_2 = 3$. Der Auswahlbereich umfasst hier die Nutzerindizes von 10 bis 14 mit der Randsumme 15.

Sollte darin kein zulässiger Sicherungsquader gefunden werden können, müssen alle drei Elementarindizes in Schleifen durchlaufen werden, wobei dann immer auch die Randsummen einzubeziehen sind, d.h. E_1 läuft von 1 bis 4, E_2 läuft von 1 bis 4 und E_3 von 1 bis 3. Das bedeutet aber noch nicht, dass damit die Gesamttabelle Auswahlbereich geworden ist, denn bezüglich der anderen Nutzerindizes mag es ja noch Eingrenzungsmöglichkeiten geben!

Der mit solch einem Verfahren zu erreichende Rechenzeitgewinn wird zwar durch Mehrfachbewertungen von Dummies zum Teil kompensiert (siehe 6.2.2.1), dennoch hat sich die „fiktive vollständige Tabelle“ bewährt (vgl.A4.2) – gerade auch im Hinblick auf den enormen Platzbedarf gespeicherter vollständiger Tabellen.

Das zuletzt vom LDS NRW entwickelte EDV-Programm QUIT (**Qu**aderverfahren **it**erativ) nutzt die temporäre Abgrenzung von Teiltabellen zur Reduzierung der Rechenzeit aus, um so auch umfangreichere Tabellen noch hinreichend sichern zu können. Testläufe haben dabei ergeben, dass die Quaderauswahlmöglichkeiten noch erheblich weiter eingegrenzt werden mussten. Ein ganz wesentlicher Rechenzeitgewinn konnte bei höher aggregierten Pivot-Elementen durch Rückverfolgung der „Spuren“ des Pivots über die zu ihm beitragenden niedriger aggregierten primär geheimen Werte erreicht werden, indem die Auswahl nur auf Quader beschränkt wurde, die alle zum Pivot beitragenden primär geheimen Werte enthalten (siehe den folgenden Abschnitt 7.2).

Nach ersten Erfahrungen mit diesem Konzept trifft man beim Aufsuchen von solchen hochdimensionalen Quadern immer wieder auf offene nicht sperrbare Tabellenwerte wie strukturelle Nullen, Tabellenwerte mit zu kleinen Schätzintervallen etc., was letzten Endes zu Sperrungen in hochaggregierten Randsummen führt. Es konnte aber inzwischen auch gezeigt werden, dass die Starrheit der Quaderstruktur zu Gunsten einer optimaleren Auswahl von Sperrkandidaten teilweise aufgegeben werden kann (vergleiche auch 6.2.2.2 zur Optimierung der Quaderauswahl durch Verschiebung von Dummies und insbesondere 7.3), ohne dabei auf Sicherheit und auf für die Quaderaus-

wahl wichtige Symmetrie-Eigenschaften der Quaderstruktur verzichten zu müssen. Solche erlaubten Quaderdeformationen sind im EDV-Programm QUIT bereits realisiert und führen durch die Reduktion der Quaderdimension auch zu Rechenzeiteinsparungen.

7.2 Rückverfolgung von Primärsperrungen

Der folgende Abschnitt behandelt die Sicherung von Primärsperrungen auf höheren Aggregationsstufen. Solche Geheimhaltungsfälle haben ihren Ursprung in den zugehörigen Primärsperrungen niedrigerer Verdichtung. In Verallgemeinerung des Problems der Auswahl von Sicherungsquadern im Falle von Einzelangaben in den Tabellenrändern (vgl. 2.1.2.2) können Quader auch zum Schutze von „gewöhnlichen“ Primärsperrungen in höherer Hierarchie so ausgewählt werden, dass sie zugehörige Primärfälle in den niedrigeren Stufen mit einschließen.

Die Sicherung höher aggregierter Pivots betrifft hier auch höher aggregierte Einzelangaben. Diese werden also nicht allein wie in Abschnitt 2.1.1 nur im Inneren einer vollständigen Tabelle als Pivots behandelt, sondern auch auf den höheren Niveaus als solche gesichert. Damit entfällt (für QUIT, nicht für GHMITER) von vorneherein das unter 2.1.2.2 beschriebene Einzelangabenproblem bei sich nicht addierenden Fallzahlen!

7.2.1 Zur Motivation: Rechenzeiteinsparung

Durch die Vorschrift, zu einem primär geheimen Tabellenwert beitragende Primärfälle niedrigerer Aggregation (nachgeordnete Primärsperrungen) in den Sicherungsquader mit einzubeziehen, lassen sich die gerade bei hochaggregierten Primärsperrungen zwangsläufig sehr großen Auswahlmöglichkeiten unter den Sicherungsquadern ganz erheblich einschränken. Eine so drastische Einschränkung von Quaderauswahlmöglichkeiten kann u.A. damit gerechtfertigt werden, dass die auf die Wertgröße abzielende Quadersummen-Optimierung nur marginal gestört wird, weil bei dieser Quaderauswahl die mit dem Pivot auf gleicher Aggregationsstufe stehenden Tabellenwerte über ihren gesamten Bereich in der Teiltabelle variiert werden. Diese Einschränkung trifft also die niedrigeren Aggregate und die haben auf die Quadersumme einen entsprechend geringeren Einfluss.

Außerdem werden im Falle von Einzelquadersicherungen durch solchermaßen ausgewählte Sicherungsquader von vorneherein alle dem Pivot nachgeordneten, von seinem Sicherungsquader erfassten Primärfälle mitgesichert: Die Werte aller zum Pivot beitragenden Primärfälle sind höchstens so groß wie der Pivot-Wert selbst und haben folglich in diesem Sicherungsquader eine hinreichend große relative Spannweite. Diese Mitsicherung von „Strangwerten“ wirkt sich ebenfalls rechenzeitsparend aus und liegt ganz im Sinne der Vermeidung von Sekundärsperrungen.

Es muss allerdings ausdrücklich darauf hingewiesen werden, dass die Mitsicherungsmöglichkeit von einem Pivot nachgeordneten Primärfällen nur dann besteht, wenn keine Doppelquadersicherung erforderlich ist. Doppelquader sichern eben stets nur das Pivot und nicht auch noch andere Quaderelemente, wie es auch das nachfolgende Beispiel in Abb. 7.4 zeigt.

Abb. 7.4

Beispieltabelle für gesichertes Pivot aber ungeschützte nachgeordnete Primärsperrungen

Wert	**Wert sekundär**	**∑ sekundär**
	Einzelwert	**Einzelwert**
Σ	**Σ**	**ΣΣ**

Wert	**Wert primär**	**∑ primär**
	Einzelwert	**Einzelwert**
Σ	**Σ**	**ΣΣ**

Σ*	**Σ***	**ΣΣ Pivot**
Σ*	**Σ***	**ΣΣ sekundär**
ΣΣ*	**ΣΣ***	**ΣΣΣ**

In dieser dreidimensionalen Tabelle sind dem Pivot (rechtes oberstes Eckfeld der untersten Tabelle) zwei weitere „gewöhnliche“ primär geheime Werte nachgeordnet (unterstrichene Werte in der mittleren Tabelle). Sie wären beide mit einem Quader für das Pivot mitgesichert, wenn nicht die Einzelangaben z.B. in der mittleren Tabelle die Sicherung des Pivots mit einem zweiten Quader erzwängen: Ein Sicherungsquader besteht aus dem Karree der geheimen Werte in der mittleren Tabelle und aus der Projektion dieses Karrees in die unterste, die Summentabelle. Der andere Sicherungsquader enthält statt der geheimen Werte der mittleren die geheimen Werte der oberen Tabelle und die entsprechenden geheimen Werte der Summentabelle.

Die dem Pivot in obiger Tabelle, Abb. 7.4, nachgeordneten (unterstrichenen) primär geheimen Werte sind nicht geschützt, sondern können von dem Melder der Einzelangaben der mittleren Tabelle berechnet werden. Der Pivot-Wert selbst lässt sich hingegen von keinem der Einzelmelder berechnen. Die nachgeordneten Primärsperrungen müssen also noch zusätzlich geschützt werden, erst dann ist die Tabelle der Abbildung 7.4 als ganze hinreichend gesichert. (Diese zusätzlichen Sperrvermerke sind nicht eingetragen, weil man ja davon ausgehen soll, dass die Tabelle schon so ausreichend gesichert ist.)

Eine Doppelquadersicherung ist beispielsweise auch vorzunehmen bei einem Pivot, das selbst nicht Einzelangabe ist, unter dessen nachgeordneten Primärsperrungen aber Einzelangaben vorkommen: Der Einzelmelder solcher nachgeordneter Einzelangaben kann mit seinem Wissen (durch Einbringen seiner Angabe) die Quadergleichungen lösen und damit den ihm fremden Pivotwert berechnen. Die Berechnung des Pivots gelingt jedoch nicht, wenn noch ein zweiter Quader für das zu schützende Pivot ausgewählt wird, der die Einzelangaben des ersten nicht ent-

hält. D.h. unter diesen Umständen ist Doppelquadersicherung angezeigt und eine Mitsicherung von Strangwerten ausgeschlossen.

Andererseits können die dem Pivot einer Einzelangabe nachgeordneten Einzelangaben - wie Einzelangaben im Rand – niemals eine Doppelquadersicherung veranlassen, denn sie betreffen alle ein und den selben Wert von ein und dem selben Berichtenden (vgl.2.1.2). Die dem Pivot einer Einzelangabe nachgeordneten Einzelangaben sind also genau wie die nachgeordneten Primärsperrungen eines mit nur einem Quader zu schützenden Pivots mit diesem Pivot bereits mitgesichert. Auch damit ist ein erheblicher Rechenzeitgewinn verbunden. – Die Strangverfolgung bei Primärsperrungen ist in gewisser Hinsicht eine Verallgemeinerung der Behandlung von Einzelangaben im Rand.

Ob eine Mitsicherung von nachgeordneten Primärsperrungen genutzt werden kann oder nicht, in jedem Fall bringt die Beschränkung auf Sicherungsquader mit Strangwerten eine drastische Verkürzung der Quaderauswahl und damit eine erhebliche Rechenzeitverkürzung. Das gilt in gleicher Weise auch für die Auswahl von Doppelquadern, wo man den selben Algorithmus wie bei der Einzelquadersuche benutzt.

7.2.2 Zur Realisation: Aufbau von Primärsperrungssträngen

Zu bearbeiten sei eine hierarchisch gegliederte, d.h. durch Zwischensummen unterteilte (Teil-)Tabelle der Dimension m, die zur Beseitigung ihrer Zwischensummen zur vollständigen Tabelle der Dimension n aufgestockt worden sei. (Zur Dimensionsaufstockung von (Teil-)Tabellen siehe vorangegangenen Abschnitt 7.1.) Das darin zu behandelnde Pivot-Element, zu dem niedriger aggregierte primär geheime Werte, Strangwerte, beitragen sollen, gehöre zu einer Untergesamtheit von gleichrangigen Tabellenwerten, d.h. von Tabellenwerten mit gleichem Aggregationsniveau.

Wegen der Übereinstimmung ihrer Aggregationsstufen mit denen des Pivots können die Werte dieser Untergesamtheit nichts zum Pivot beitragen; sie sind daher bei der Quaderauswahl mit Fixierung durch Strangwerte noch frei verfügbar. Hier werde zunächst eine Tabelle frei von sperrbaren Dummies betrachtet. Die Dimension m´ der Untergesamtheit mit dem Pivot gleichrangiger Tabellenwerte ergibt sich dann aus der Dimension n der aufgestockten (Teil-)Tabelle abzüglich der Anzahl nachgeordneter Primärsperrungen.

Es ist $0 \leq m' \leq n - 1$, weil mindestens ein primär geheimer Wert dem Pivot nachgeordnet sein soll. Daher ist diese m´-dimensionale Untergesamtheit im Allgemeinen mit keiner der Untertabellen der hierarchisch gegliederten Tabelle identisch, weil die Untertabellendimension immer mit der Dimension der Gesamttabelle m übereinstimmt, wogegen m´ größer und auch kleiner als m sein kann.

m´ ist Null, wenn das Pivot die größte Eckfeldsumme ist. m´ stimmt mit der Dimension m der „unaufgestockten“ (Teil-)Tabelle überein, wenn das Pivot ganz im Innern der Untertabelle höchster Aggregation der „unaufgestockten“ (Teil-)Tabelle liegt und m´ ist größer als m, wenn sich das Pivot im Innern einer Untertabelle niedrigerer Aggregation befindet.

Das Pivot der Beispieltabelle von Abb. 7.4 gehört zu einer Untergesamtheit der Dimension $m' = 1$, d.h. nur ein Elementarindex kann noch unter Beibehaltung aller Aggregationsstufen des Pivots variieren, der Zeilenindex. Der

primär geheime Wert im Rand der mittleren Tabelle, als Pivot betrachtet, liegt im Inneren der Untertabelle höchster Aggregation. Die Dimension dieser Untergesamtheit ist folglich die der „unaufgestockten“ Tabelle, d.h. $m' = m = 2$. Eine erste Einschätzung des an m' zu messenden Umfangs an Quaderauswahlmöglichkeiten ergibt sich aus nachfolgender Beschreibung der Einbeziehung von nachgeordneten Primärsperrungen in den Sicherungsquader eines Pivots.

Ausgehend von den n gegebenen Elementarindizes eines in seiner (Teil-)Tabelle zu sichernden Pivots höchster Verdichtung wird zunächst in den Elementarindex-Segmenten zweithöchster Aggregationstufe nach einem weiteren, zu einer anderen Primärsperrung gehörigen Index gesucht, wobei die $n - 1$ anderen Elementarindizes als Pivotindizes fest bleiben. Zur zweithöchsten Aggregationsstufe existiert bezüglich eines der Elementarindex-Segmente immer eine zum Pivot beitragende andere Primärsperrung, weil das Pivot in einer um Eins höheren Aggregationsstufe real vorliegt.

Dagegen findet sich unter den Elementarindizes der Segmente gleichen Niveaus wie das Pivot keine Primärsperrung, die zum betrachteten Pivot beitragen könnte. Daher beginnt auch die Suche nach Strangelementen erst in einem der zweithöchsten Elementarindex-Segmente. Der in diesem ersten Schritt gefundene Elementarindex ist von dem des Pivots verschieden, kann daher als einer der Diametralindizes angenommen und als solcher gemerkt werden.

Im nächsten Schritt wird unter den dritthöchsten Elementarindizes nach einem Index eines vom Pivot verschiedenen primär geheimen Wertes gesucht, wobei die noch verbliebenen anderen n-2 Pivotindizes und der im vorhergehenden Schritt gefundene Diametralindex fest bleiben. Wieder wird der neu gefundene, vom Pivotindex verschiedene Index als ein weiterer Diametralindex gemerkt. Auf diese Weise fährt man fort, bis das unterste Aggregationsniveau erreicht ist.

Wenn man dabei auf keinen Dummy (z.B. als „Hinterland“ einer kreisfreien Stadt) gestoßen ist, so sind bei s nachgeordneten Primärsperrungen bereits $s = n - m'$ Diametralindizes gefunden (s steht für Strang). Im Dummy-Fall gibt es noch die Auswahl innerhalb der d Elementarindex-Gliederungen mit Dummy-Werten zu den „Hinterlandgemeinden“, die zweckmäßig zusammen mit den noch ausstehenden m' Elementarindizes höchster Aggregationsstufe auf der Suche nach geeigneten Indizes (Quaderauswahl-Kriterien) zu durchlaufen sind.

Was nun noch fehlt, sind die m' Diametralindizes der jeweils höchsten Aggregationsstufen, die oben noch nicht mit Hilfe von zum Pivot beitragenden Primärfällen niedrigerer Aggregation festgelegt werden konnten und die nun zusammen mit den d Dummy-Elementarindizes abgetastet werden müssen. Bezüglich jedes dieser Indizes besteht u.A. die Vorschrift, die im Sinne der kleinsten Quadersumme günstigste Quaderauswahl zu treffen, wobei gerade zwischen den höchsten Aggregaten differenziert wird. Das kommt einer optimalen Quaderauswahl besonders zugute.

Auf diese Weise hat man bereits alle gesuchten n Diametralindexwerte mit $n = s + m' + d$ gefunden, ohne dabei den ganzen, durch das Aggregationsniveau des Pivotelements festgelegten hochdimensionalen Raum abtasten zu müssen, denn s Indizes sind davon bereits durch die s nachgeordneten Primärsperrungen festgelegt.

Wie bemerkt, werden dabei nur die m´ Elementarindizes zur höchsten Aggregation und ggf. auch noch erlaubte Dummy-Elementarindizes nach einem geeigneten Quader durchsucht, was oft – abgesehen von den Dummies - genau der Suche nach einem Sicherungsquader in einer einzigen Untertabelle höchster Verdichtung gleichkommt, wenn nämlich m´ = m gilt, oder was im Falle m´ < m sogar noch stringenter ist. Der aus diesem Vorgehen resultierende Rechenzeitgewinn ist damit unmittelbar evident.

7.2.3 Zur technischen Durchführung: Strang-Aufbau eines Pivot-Elements

Das oben skizzierte Vorgehen zur Einbeziehung eines ganzen Stranges von zu einem Pivot beitragenden primär geheimen Tabellenwerten in einen Sicherungsquader sei zunächst anhand der Beispieltabelle, Abb. 7.5 erläutert.

Abb. 7.5

Beispieltabelle zur Strangverfolgung von Primärsperrungen über mehrere Hierarchieebenen

		1			2			3			4		
		1	2	3	1	2	3	1	2	3	1	2	3
1	1	0	W	W	W	0	W	W	W	W	*D	*D	W
	2	W	W	W	W	0	W	W	0	W	*D	*D	W
	3	0	0	0	W	W	W	0	W	W	*D	*D	W
	4	W	W	W	W	W	W	W	W	W	*D	*D	W
2	1	W	0	W	W	W	W	W	W	W	*D	*D	W
	2	0	0	0	0	W	W	0	W	W	*D	*D	W
	3	W	E	W	W	0	W	0	W	W	*D	*D	W
	4	W	E	P	W	W	W	W	W	W	*D	*D	P
3	1	*D	*D	*D	*D	*D	*D	*D	*D	*D	*D	*D	*D
	2	*D	*D	*D	*D	*D	*D	*D	*D	*D	*D	*D	*D
	3	*D	*D	*D	*D	*D	*D	*D	*D	*D	*D	*D	*D
	4	W	W	W	W	W	W	W	W	W	*D	*D	W

Diese Beispieltabelle ist eine nach zwei Merkmalen hierarchisch gegliederte Teiltabelle, die zu einer fünfdimensionalen vollständigen Tabelle aufgestockt wurde. Bezüglich der Spaltengliederung ist nur ein Elementarindex-Wert dritter, höchster Aggregation dargestellt; die Zeilengliederung wurde vollständig erfasst. Das zu sichernde Pivot befindet sich im zweiten Zeilen- und im vierten Spalten-Block der besonders abgesetzten Untertabellen-Blöcke. Es nimmt in dieser Untertabelle das Summen-Eckfeld ein. In den Tabellenfeldern sind nur die Werte als Platzhalter ausgewiesen: W = Wert, P = Primärsperrung, E = Einzelangabe und *D = Sternchensumme.

Die Elementarindizes des Pivot-Elements sind demnach (2; 4; 1; 4; 3). Zu diesem Pivot tragen die Primärfälle (nach absteigendem Aggregationsniveau) (2; 4; 1; **1**; 3) als „gewöhnliche“ Primärsperrung, (2; 4; 1; **1**; **2**) als Einzelangabe und (2; **3**; 1; **1**; **2**) ebenfalls als Einzelangabe bei. - Hier liegt also der oben angesprochene Fall vor, der eine Doppelquadersicherung erzwingt, obwohl die betreffenden Einzelangaben selbst zum Pivot beitragen! - Als Diametralindizes leiten sich aus diesem Strang die folgenden Indizes ab:

Ausgehend vom Pivot (2; 4; 1; 4; 3) findet man für die zweite (durch Unterstrich markierte) Aggregationsstufe, die zweithöchste Verdichtung des Spaltenindex, als benachbartes Quaderelement die Primärsperrung (2; 4; 1; **1**; 3). Die zweite Verdichtungsstufe ist beim Zeilenindex bereits die höchste und kommt daher nicht in Betracht. Ausgehend von dieser Primärsperrung (2; 4; 1; **1**; 3) findet man für die beiden Elementarindizes 4 und 3 erster Aggregationsstufe die Primärsperrung (2; 4; 1; **1**; **2**). Beim Index 4 wird man hier noch nicht fündig; das gelingt erst nach Variation des vierten Elementarindex der zuletzt gefundenen Primärsperrung (2; 4; 1; **1**; **2**); man findet für die unterste Aggregation (2; **3**; 1; **1**; **2**).

Durch die sukzessive Ersetzung der bei jedem Schritt unterstrichenen Pivot-Indizes durch die fettgedruckten Elementarindizes erhält man die s = n - m´ = 5 - 2 = 3 Diametralindizes (...; 3; ...; 1; 2). Die beiden durch Punkte gekennzeichneten Elementarindizes, die aus den unterstrichenen beiden im letzten Primärfall hervorgehen, müssen unter Berücksichtigung der Quaderauswahlkriterien gemeinsam „abgetastet“ werden.

Man kann die Beispieltabelle, Abb.7.5, auch als aufgestockte vierdimensionale (Teil-)Tabelle betrachten. Dann ist das Pivot mit festgehaltenem eingeklammerten dritten Elementarindex (2; 4; (1); 4; 3) ein Randsummenelement, dessen dritter Index bei der Quaderauswahl nicht mehr variiert werden kann, obwohl die Strangwerte immer noch die selben sind wie in obigem Beispiel. Hier hat man es offenbar mit der gleichen Situation zu tun wie beim ersten Beispiel der Abbildung 7.4 mit Pivot im höchsten Rand. Die Dimension der Untergesamtheit mit dem Pivot gleichrangiger Tabellenwerte ist auch in diesem Fall nur m´= 1.

Diametralwertbestimmung (formal)

a) Festlegung des Elementarindex-n-Tupels des Pivots, wobei die einzelnen Elementarindizes in der Reihenfolge absteigender Verdichtung zu sortieren sind. Zusammenfassen aller Elementarindizes höchster Aggregationsstufen, dann der Elementarindizes zweithöchster Aggregationsstufen und weiter nach absteigenden Aggregationsstufen. Dadurch entstehen innerhalb des Pivot-n-Tupels Index-Gruppen gleichhoher Aggregationsstufen.

b) Beginnend mit den Indizes zweithöchster Aggregation, Abarbeiten der Indexgruppen nach absteigenden Aggregationsstufen, wobei ein Index der gerade zu behandelnden Indexgruppe nach dem anderen innerhalb seines Segments solange variiert wird, bis ein anderer eine Primärsperrung markierender Elementarindex gefunden wurde. Die so aufgesuchten s = n - m´ - d neuen Elementarindizes werden als Indizes eines zum gegebenen Pivot-Element diametralen Wertes abgespeichert. Die mit d gekennzeichnete Anzahl von Segmenten mit erlaubten Dummies (Hinterland, siehe 7.2.2) wird in c) mitbehandelt.

c) Auswahl der nun noch offenen m´ Elementarindizes höchster Verdichtung zuzüglich der noch unbestimmten d Dummy-Indizes durch Variation dieser Indizes jeweils innerhalb ihrer Segmente, wobei die Pivot-Indizes selbst auszunehmen sind (Diametralität). Um hier die richtige Auswahl zu treffen, muss zu jedem Diametralindex-n-Tupel der zugehörige n-dimensionale Quader aufgesucht und nach den vorgegebenen Auswahlkriterien beurteilt werden.

d) Beim Aufsuchen der zum Pivot beitragenden primär geheimen Tabellenwerte kann eine Aufspaltung in Nebenstränge eintreten, wenn nämlich in einem Segment mehr als ein weiterer Primärfall indiziert ist. In solch einem Fall wird der erste gefundene Index gewählt und der Strang von hier aus weiterverfolgt. Die weiteren Primärfälle werden dann bei der Sicherung der nicht in den Strang einbezogenen Primärsperrungen behandelt.

Es bleibt nun noch die Frage, ob denn mit obigem Vorgehen alle angesprochenen nachgeordneten primär geheimen Werte (Strangwerte) tatsächlich auch in den Sicherungsquader des betrachteten Pivots einbezogen worden sind. Zur Klärung wird nochmals das Pivot in der Beispieltabelle 7.5 und seine Strangwerte betrachtet.

geheimer zu schützender Wert (Pivot):	(2, 4, 1, $\underline{4}$, 3)	= (g_1 ,g_2 ,g_3 ,g_4 ,g_5)
zuerst gefundener Strangwert:	(2, 4, 1, **1**, $\underline{3}$)	= (g_1 ,g_2 ,g_3 ,$\mathbf{d}_4$,g_5)
zweiter gefundener Strangwert:	(2, $\underline{4}$, 1, **1**, **2**)	= (g_1 ,g_2 ,g_3 ,$\mathbf{d}_4$,$\mathbf{d}_5$)
dritter gefundener Strangwert:	($\underline{2}$, **3**, $\underline{1}$, **1**, **2**)	= (g_1 ,$\mathbf{d}_2$,g_3 ,$\mathbf{d}_4$,$\mathbf{d}_5$)
gesetzter Diametralwert:	(**1**, **3**, **2**, **1**, **2**)	= ($\mathbf{d}_1$,$\mathbf{d}_2$,$\mathbf{d}_3$,$\mathbf{d}_4$,$\mathbf{d}_5$)

Mit den Indizes seiner beiden zueinander diametralen Werte (g_1 ,g_2 ,g_3 ,g_4 ,g_5) und ($\mathbf{d}_1$,$\mathbf{d}_2$,$\mathbf{d}_3$,$\mathbf{d}_4$,$\mathbf{d}_5$) schließt der Sicherungsquader des zu schützenden Pivots alle Strangwerte ein, weil jeder Strangwert ausschließlich durch die Indizes dieser beiden zueinander diametralen Werte indiziert ist (Quaderdefinition, 2.1.1).

Um diese Aussage zu verallgemeinern, ist vom Index-n-Tupel (g_1 ,g_2 , ... , g_i , ... , g_n) eines zu schützenden Pivots in einer n-dimensionalen Tabelle auszugehen. Darin wird in der Abfolge der Strangwerte nacheinander ein g_i nach dem anderen durch ein entsprechendes, die Lage des betreffenden Strangwertes bezüglich des i-ten Elementarindex festlegenden Index als Diametralindex d_i ersetzt.

Da in diesem Prozess immer das vorangegangene Index-n-Tupel bis auf den im jeweiligen Schritt der Strangverfolgung umgesetzten Index beibehalten wird und der gerade umgesetzte Index auch als Diametralindex zum entsprechenden Pivot-Index angenommen wird, enthält jeder Strangwert nur Indizes der beiden zueinander diametralen Werte. D.h. jeder Strangwert gehört zu einem Sicherungsquader, bei dem der zum Pivot diametrale Quaderwert durch eben diese umgesetzten Indizes d_i indiziert ist. Diese Aussage ist ganz offensichtlich unabhängig von den nach Abarbeitung des Stranges noch frei wählbaren $m' + d$ Elementarindizes zur Vervollständigung des Index-n-Tupels des Diametralwertes.

7.3 Quaderdeformation

Der folgende Abschnitt greift die unter 6.2.2.2 angesprochene Idee wieder auf, durch Versetzen von Quaderteilen Randsperrungen zu vermeiden: Da Sternchensummen nicht veröffentlicht werden, kann man einzelne Tabellenteile durch Veränderung der Anordnung von Dummy-Feldern gegeneinander verschieben, wenn alle realen Tabellensummen dabei unverändert bleiben.

7.3.1 Ausgangssituation

In hierarchisch fein gegliederten Tabellen trifft man beim Aufsuchen von vollständigen Sicherungsquadern (Definition unter 3.2.3.3) in deren hochdimensionalen (Teil-)Tabellen häufig auf verbotene Dummy-Werte oder auch auf andere nicht sperrbare Tabellenwerte, wie z. B. strukturelle Nullen oder Tabellenwerte mit verschwindenden Schätzfehlern (als bekannt vorauszusetzende Tabellenwerte). Solche Quader sind als Sicherungsquader abzulehnen. Dies führt dann in aller Regel zur Auswahl von Sicherungsquadern mit Randsummenwerten und damit zur weiteren Aufstockung der jeweiligen Teiltabelle mit der Folge weiterer zusätzlicher Sekundärsperrungen (zur Erweiterung der aufgestockten Teiltabelle vgl. 7.1).

Sternchensummen in Verbindung mit verschiebbaren Dummy-Feldern bieten nun aber eine Ausweichmöglichkeit, nämlich den Übergang von einem im Quader vorliegenden nicht sperrbaren Wert zu benachbarten, noch nicht zum Quader gehörigen sperrbaren Tabellenwerten. Unter günstigen Umständen lässt sich der gefundene inakzeptable Quader mit solchen benachbarten Werten so ergänzen, dass er als Sicherungsquader akzeptabel wird. Dabei ist das nachstehend beschriebene schrittweise Vorgehen angezeigt, das zwangsläufig zu einer gewissen Deformation des ursprünglichen Quaders führt.

Mit diesem Vorgehen verbindet sich eine Abkehr von der starren Quaderstruktur (Dekadenz der Quaderstruktur), allerdings ohne dabei auf die Vorteile der ursprünglichen Quadersymmetrie ganz verzichten zu müssen. Die Quaderauswahl erfolgt nach wie vor durch Aufsuchen eines zum zu sichernden Pivot diametralen Wertes. Auch die darauf aufbauende vorläufige Festlegung der Gesamtheit aller Quaderelemente und deren Aufteilung in die beiden zueinander komplementären Quaderteilgesamtheiten bleibt von o.g. Quader- Deformation unberührt, weil die Zugehörigkeit der Quaderelemente zu ihren Teilgesamtheiten bei der Deformation einfach auf die neuen Werte übertragen werden kann.

7.3.2 Aufsuchen von Nachbarwerten eines nicht sperrbaren Quaderwerts

In einer aufgestockten n-dimensionalen (Teil-)Tabelle sind jedem ganz im Inneren gelegenen Tabellenwert jeweils 2n Tabellenwerte als (nächste) Nachbarn zuzuordnen, zwei Werte pro Elementarindex. – Die Bezeichnung „ ganz im Inneren einer Tabelle“ für die Gesamtheit aller Tabellenwerte mit 2n Nachbarn ist praktisch eine Verschärfung der bisher benutzten Definition für das Tabelleninnere, bei dem nur die Tabellenwerte ohne die Randsummenwerte gemeint sind.

Die Gesamtheit von Nachbarn reduziert sich oft schon aus rein geometrischen Gründen: Liegt der zu ersetzende Quaderwert in der Tabelle am Rand bezüglich einer oder mehrerer Elementarindizes, so entfallen alle Nachbarn, die nur noch jenseits des Randes Platz fänden; d.h. wenn der Elementarindex des zu ersetzenden Wertes der größte oder der kleinste Indexwert ist, existiert bezüglich dieses Elementarindex´ nur ein Nachbar.

Ein anderer Grund für den Ausschluss eines Nachbarwertes als Ersatzwert für einen nicht sperrbaren Quaderwert liegt vor, wenn der Nachbar selbst schon zum vorgegebenen Quader gehört. Diese Situation tritt beispielsweise auf, wenn die Werte des gegebenen Quaders bereits nahe benachbart sind, wie es in der Beispieltabelle Abb. 7.6 für die schattierten Quaderfelder zutrifft. Diese Felder bilden auch den Ausgangsquader von dem aus ein defor-

mierter Quader aufzubauen ist, weil in der oberen Zeile seines linken oberen Karrees zwei nicht sperrbare Nullen stehen.

Die Gesamtheit der geometrisch zu akzeptierenden Nachbarn erhält man demnach auf folgende Weise:

Nacheinander wird nur ein Elementarindexwert des auszuwechselnden Quaderwertes nach dem anderen durch einen „Nachbarindex“ ersetzt, wobei die anderen Elementarindexwerte als die des auszuwechselnden Quaderwertes erhalten bleiben. Dazu wird einmal der oberhalb des zu ersetzenden Quaderwert-Index´ liegende Elementar-Indexwert genommen und einmal der darunter liegende Indexwert, sofern diese Indexwerte überhaupt existieren. Wohlbemerkt behalten dabei alle n -1 anderen Elementarindizes die Indexwerte des zu ersetzenden Quaderwertes bei. Ein Indexwert existiert (als Nachbarwert-Index), wenn er zum Segment des betreffenden Elementarindex gehört und (noch) nicht selbst Quaderindex ist. Als Beispiel siehe Abb. 7.6:

Abb. 7.6

Testtabelle zur Quadermodifikation durch Umstellung von Dummy-Feldern

		1	1 2	3	1	2 2	3	1	3 2	3	1	4 2	3
1	1	0	W	W	W	0	W	W	W	W	*D	*D	W
	2	W	W*	W*	W**	0*	W*	W	0	W	*D	*D	W
	3	0	0	0	W	W	W	0	W	W	*D	*D	W
	4	W	W	W	W**	W	W	W	W	W	*D	*D	W
2	1	W	0	W	W	W	W	W	W	W	*D	*D	W
	2	0	0	0	0	W	W	0	W	W	*D	*D	W
	3	W	E	W	W**	0	W	0	W	W	*D	*D	W
	4	W	E	P	W**	W	W	W	W	W	*D	*D	W
3	1	*D	*D	*D	*D	*D	*D	*D	*D	*D	*D	*D	*D
	2	*D	*D	*D	*D	*D	*D	*D	*D	*D	*D	*D	*D
	3	*D	*D	*D	*D	*D	*D	*D	*D	*D	*D	*D	*D
	4	W	W	W	W	W	W	W	W	W	*D	*D	W

* Umstellungsversuch 1 ** Umstellungsversuch 2 schattiert eingefärbt : Ausgangsquader

Die Darstellung zeigt eine gegebene hierarchisch gegliederte zweidimensionale Tabelle, die bezüglich beider Gliederungen aufgestockt wurde. Das Ergebnis dieser Aufstockung ist eine vierdimensionale Tabelle, die hier in Matrixform dargestellt ist (weitere Erläuterungen zu dieser Darstellungsform findet man unter 6.2.2.2). Die Elementarindizes werden von oben nach unten und von links nach rechts aufsteigend durch Nummern bezeichnet. Die in den beiden gegebenen Gliederungen höchsten Elementarindizes kennzeichnen die besonders abgesetzten Untertabellen unterster Aggregation, die niedrigsten legen die Zeilen bzw. die Spalten fest.

Der zur Sicherung des mit P markierten Pivot-Elements ursprünglich ausgewählte Quader umfasst 16 Tabellenfelder, die durch Schattierung besonders herausgehoben sind. Der auszutauschende Quaderwert in der dritten Zeile und der zweiten Spalte betrifft eine nicht sperrbare Null. Dieser Wert ist im hier vorliegenden Fall einer aufgestockten Tabelle mit der Dimension $n = 4$ durch das Elementarindex-Quadrupel (1, 3, 1, 2) festgelegt.

Daraus gewinnt man mit obiger „Regel“ die Index-Quadrupel folgender Nachbarwerte: (1, **2**, 1, 2), (1, 3, 1, **1**). Die anderen 6 Index-Quadrupel kommen nach obigen Ausführungen nicht in Frage, weil zwei von ihnen bezüglich der beiden obersten Elementarindizes außerhalb der Tabelle lägen und vier bereits selbst Quaderwerte sind.

.

7.3.3 Auswahl geeigneter Ersatzwerte aus der Nachbarngesamtheit

Um zu entscheiden, welcher der Nachbarwerte für eine Ersetzung des nicht sperrbaren Quaderwertes tatsächlich in Frage kommt, muss man berücksichtigen, dass Werteverschiebungen gegeneinander immer nur senkrecht zu einer Elementarindex-Achse mit Sternchensummenrand erfolgen darf, wo also der Achsen-Index als Summierungsindex immer auf eine Sternchensumme führt. Die anderen „Normalen-Richtungen“ (mit realem Summenrand) sind für dazu senkrechte Verschiebungen tabu, weil anderenfalls Verfälschungen in der Veröffentlichungstabelle aufträten.

Man muss demnach zwischen Elementarindizes mit Sternchensummenrand – hier kurz als Sternchenindizes bezeichnet - und Indizes mit realem Summen-Rand unterscheiden. Sternchenindizes werden daher durch ein Sternchen markiert und wie folgt definiert:

Definition:

Sternchenindizes sind Elementarindizes einer zur vollständigen Tabelle aufgestockten ursprünglich hierarchisch gegliederten (Teil-)Tabelle, die als Summierungsindizes zu nicht zu veröffentlichenden Sternchensummen führen.

Nach dieser Definition hängt die Eigenschaft, Sternchenindex zu sein, von den Ausprägungen der anderen n-1 Elementarindizes ab! D.h. ein gerade als Summationsindex fungierender Elementarindex kann einmal Sternchenindex sein und er kann, nach Verändern eines anderen Elementarindex, also beim Übergang zu einer anderen Randsummenwert-Berechnung (aber immer noch mit dem selben Elementarindex als Summationsindex), auch als Elementarindex ohne Sternchenmarkierung in Erscheinung treten. Das hängt damit zusammen, dass jeder Tabel-

lenteil mit Summenumrandung immer auch Summen ohne Sternchen enthält – anderenfalls wäre die Sternchensummbildung entbehrlich -. (Zur Tabellenvervollständigung durch Dimensionsaufstockung siehe 7.1.)

Um aus der gemäß 7.3.2 bereits vorliegenden Nachbarngesamtheit die für die Ersetzung des nicht sperrbaren Quaderwertes geeigneten Nachbarwerte auszuwählen, markiert man zunächst die Sternchenindizes aller Quaderwerte und die Sternchenindizes der o.g. „geometrischen“ Nachbarn. Für die Ersetzung kommen nur noch diejenigen Nachbarn in Frage, die mit dem zu ersetzenden Quaderwert einen Sternchenindex als Sternchen-Normalenindex (der Hyperebene, in der verschoben wird) gemeinsam haben und die durch Verändern eines anderen Elementarindex des zu ersetzenden Quaderwertes entstehen.

Auswahl von Nachbarn als Ersatzwerte:

a) Markieren aller Sternchenindizes der Elementarindex-n-Tupel aller Quaderwerte des zu deformierenden Quaders gemäß o.g. Definition.

b) Umsetzen eines Elementarindex des auszutauschenden Quaderwertes in seinen Nachbarindex, den Verschiebeindex, durch Addition oder Subtraktion von Eins, sofern dies nicht aus dem Segment hinausführt.

c) Durch Vergleich des mittels b) erhaltenen Indexwertes mit dem nicht zu ersetzenden Quaderindex zum gleichen Segment (es gibt ja immer zwei Quaderindizes in jedem Segment, den zu ersetzenden und einen weiteren) prüfen, ob der gefundene Nachbar selbst als Sperrpartner verboten oder bereits Quaderwert ist. Falls ja, verwerfe diesen Index und wähle ggf. vermöge b) einen neuen Nachbarindex aus.

d) Markieren aller Sternchenindizes der mit b) und c) gefundenen Nachbarwerte. Als Ersatzwerte des nicht sperrbaren Quaderwerts werden diejenigen sperrbaren Nachbarn vorgesehen, die mit dem zu ersetzenden Quaderwert einen Sternchenindex als Sternchen-Normalenindex der Hyper-Ebene, in der verschoben werden soll, gemeinsam haben.

Für unseren Beispiel-Quader in der Abbildung 7.6 ergibt sich für das Elementarindex-Quadrupel des auszuwechselnden Quaderwertes nach der Sternchenmarkierung gemäß a) (1*, 3, 1*, 2). Addition von Eins zu nur einem der vier Elementarindizes gemäß b) führt gemäß c) in allen vier Fällen zu einem Quaderindex zum selben Segment, kommt also als Nachbar, zu dem übergegangen werden soll, nicht in Frage. Als Ersatzwerte stehen daher nur noch die durch (1*, 2, 1*, 2) und (1*, 3, 1*, 1) indizierten Nachbarn zur Verfügung, die zugleich auch der o.g. Vorschrift d) gerecht werden.

7.3.4 Aufsuchen mitzuverschiebender Quaderwerte

Ein einzelner Quaderwert kann in der Regel nicht alleine gegen andere Quaderwerte „verschoben“ werden, weil damit eine Verschiebung gegenüber einer zu veröffentlichenden Summe verbunden wäre. Die Verfälschung von Veröffentlichungssummen lässt sich jedoch vermeiden, wenn die betroffenen Veröffentlichungs-Summen mit allen

zu ihnen beitragenden Tabellenwerten mitverschoben werden. Dies geschieht durch Einfügung von Dummy-Feldern und wird nach der Rückführung in die Veröffentlichungstabelle ohnehin unsichtbar; lediglich die Sperreintragungen aufgrund von Quaderdeformationen lassen noch solche Verschiebungen erahnen.

Das bedeutet für den Übergang zu einem gemäß 7.3.3 ausgewählten Nachbarwert, dass zusammen mit dem auszutauschenden Quaderwert selbst noch alle anderen, zur selben n-2-dimensionalen „Quaderebene“ gehörigen Quaderwerte mitverschoben werden müssen, die durch den Verschiebeindex zum ausgewählten Nachbarn und den gemeinsamen Sternchenindex definiert ist. Nur dadurch, dass die anderen Quaderindizes bei der Verschiebung dieser Quaderebene erhalten bleiben und die Quaderwerte dieser Ebene alle den selben Normalenindex (zu dem senkrecht verschoben wird) als Sternchenindex gemeinsam haben, wird ein Erhalt von realen Quadersummen überhaupt erst garantiert. Daraus leitet sich eine Vorschrift für eine erlaubte Quaderverformung ab.

Bestimmung der Gesamtheit zu verschiebender Quaderwerte:

Zu gegebenem Nachbarwert findet man die Gesamtheit zu verschiebender Quaderwerte, indem man aus der Quadergesamtheit alle Index-n-Tupel auswählt, die einen fest vorgegebenen Sternchenindex als Sternchen-Normalenindex (der die Hyperebene markiert, in der verschoben wird) und den zu ersetzenden Elementarindex des nicht sperrbaren Quaderwertes gemeinsam haben. Dann trägt man bei all diesen Index-n-Tupeln für den zu ersetzenden Quaderwert-Index den Nachbar-Index (Verschiebeindex) ein, falls der Sternchennormalenindex in keinem dieser neuen n-Tupel zu einem Nicht-Sternchenindex konvertiert, sonst muss die selbe Prozedur mit einem anderen Sternchennormalenindex oder gar mit einem anderen Nachbarn durchgeführt werden.

Für den Ersetzungsfall bei unserem Quader in der Beispieltabelle, Abb. 7.6, kann man gemäß 7.3.3 als Nachbarwert (1*, 2, 1*, 2) wählen. Als Hyperebene, in der verschoben werden soll, kommt gemäß obiger Vorschrift z.B. die durch den ersten Sternchenindex gekennzeichnete Gesamtheit von Tabellenwerten, (1*,...,...,...), in Frage. Die zu verschiebende n-2-dimensionale Quaderebene ergibt sich mit der zusätzlichen Einschränkung, dass der zweite Index, von 3 nach 2 verschoben werden soll. Es sind demnach alle Quaderwerte (1*, 3,...,...) nach (1*, 2,...,...) zu verschieben. Dazu wählt man alle Quaderwerte mit den ersten beiden Indizes 1* und 3 aus und ersetzt in deren Index-Quadrupeln den zweiten Index 3 durch 2: (1*, 2, 1*, 2), (1*, 2, 1, 3), (1*, 2, 2*, 2), (1*, 2, 2, 3).

Unter den so indizierten Werten findet sich in unserem Beispielquader wieder eine nicht sperrbare Null (horizontal schraffiert), so dass obiger Vorgang ggf. wiederholt werden muss. Die obere Grenze der Anzahl möglicher Wiederholungen des o.g. Deformationsvorgangs bei einem einzelnen Quader wird sich nur anhand von Auswertungen von Realdaten experimentell optimieren lassen. Der hier beschriebene Verschiebe-Algorithmus stellt, wenn man auch Mehrfachanwendungen auf den selben Sicherungsquader in Betracht zieht, eine neue „innere“ Iteration dar, die offenbar an die Stelle der Untertabelleniteration früherer Geheimhaltungsverfahren tritt.

7.3.5 Begründung für die Quaderdeformation

7.3.5.1 Einzelwertsicherung bei partieller Aufstockung

Bei der hier zu diskutierenden Einzelwertsicherung kann man sich den gesamten Sicherungsprozess einer hierarchisch gegliederten Tabelle grob betrachtet folgendermaßen vorstellen: Ein zu schützender geheimer Wert wird durch eine geeignete Gesamtheit gesperrter Tabellenwerte gesichert, dadurch entsteht eine neue Tabelle. Damit beginnt dann der ganze Sicherungsprozess wieder von vorne. Aus der neuen hierarchisch gegliederte Tabelle sucht das Programm einen anderen zu schützenden geheimen Wert heraus, sichert ihn beispielsweise wieder mit einem geeignet deformierten Quader gesperrter oder noch zu sperrender Werte und fährt so fort, bis jeder einzelne zu schützende Tabellenwert gesichert ist.

Dabei bleiben die in den vorangegangenen Schritten des Sicherungsprozesses eingefügten Dummies unberücksichtigt, weil jede aktuelle Sicherung eines geheimen Wertes zunächst mit dem Aufsuchen des Pivots in der jeweils neuesten unaufgestockten hierarchisch gegliederten Tabelle beginnt, die noch keine Dummies enthält. Erst nach dem Auffinden des zu bearbeitenden Pivots kann die Abgrenzung einer durch dieses Pivot-Element bestimmten Teiltabelle erfolgen, die dann ggf. unter Eintragung von Dummies – mit für diese Teiltabelle ganz spezifischen Positionen - temporär aufgestockt wird (vgl. partielle Aufstockung, 6.2.2.3). Im aktuellen Prozessschritt haben die Dummy-Positionen vorangegangener Sicherungen somit keinerlei Einfluss auf die bei der Quaderauswahl gerade wirksame Geometrie der speziellen dem Pivot zugeordneten Teiltabelle.

Um nun sicherzustellen, dass ein beliebig herausgegriffenes Pivot-Element durch eine gegebene Gesamtheit gesperrter Werte in seiner Veröffentlichungstabelle hinreichend geschützt ist, genügt es, diese Schutzgesamtheit losgelöst von allen anderen Sperrungen zu überprüfen (vgl. 3.1.2.3). D.h. bei der Überprüfung darf man ohne Gefährdung der Geheimhaltung annehmen, dass alle nicht zu dieser Schutzgesamtheit gehörenden Tabellenwerte, die hier (unabhängig von der Struktur der Schutzgesamtheit) kurz als Quaderaußenwerte bezeichnet werden sollen, offen seien.

Mit dieser Auffassung von sekundärer Geheimhaltung lässt sich überhaupt erst eine schrittweise Einzelwertsicherung rechtfertigen. Das gilt insbesondere für eine Einzelwertsicherung nach obigem Muster, wobei auch noch die jedem Pivotwert zugeordnete Teiltabelle ganz individuell mit Dummies erweitert werden kann. Ist demnach das Pivot erst einmal mit einer Schutzgesamtheit, z.B. einem Quader, im Inneren einer Teiltabelle hinreichend gesichert, so auch in der hierarchisch gegliederten Veröffentlichungstabelle, wie immer die anderen geheimen Werte dort geschützt wurden.

Darüber hinaus führt die Behandlung der Quaderaußenwerte als offene Tabellenwerte bei der Sicherheitsüberprüfung des Pivots noch zu einer wesentlichen Reduktion der obigen (Teil-)Tabelle: Alle Tabellen-Hyperebenen, die ausschließlich Quaderaußenwerte enthalten, können mit all ihren Randsummen von vorneherein von den zugehörigen Tabellen-Summen-Hyperebenen subtrahiert werden, so dass sie aus der aufgestockten Tabelle gänzlich verschwinden.

7.3.5.2 Sicherung mit deformiertem Quader

Hier wird ein Tabellenschnitt mit nur zwei Elementargliederungen als Repräsentant der aufgestockten n-dimensionalen Tabelle betrachtet. Er lässt sich durch die o.g. Subtraktion aller Quaderaußenwerte auf eine 2x2-Tabelle mit Zeilen- und Spaltenrandsummen zurückführen (Abb. 7.7). Jedes Tabellenfeld repräsentiert darin eine (n-2)-dimensionale Tabelle. Betrachtet werden jeweils nur die nulldimensionalen Pendants der (n-2)-dimensionalen „Feld-Tabellen", die in die betreffende 2-dimensionale Tabelle fallen. Die Abb. 7.7 stellt zwar hinsichtlich ihres Sperrmusters nur einen Spezialfall dar, alle anderen Fälle sind aber nach der Diskussion dieses Falles unmittelbar evident.

Abb. 7.7: **Zweidimensionale Spur des undeformierten Sicherungsquaders**

Pivot + ε	**Dummy**	**$\Sigma_{geheim} + \varepsilon$**
$W_{geheim} - \varepsilon$	**$Wert_{verboten}$**	**$\Sigma_{geheim} - \varepsilon$**
Σ*	**Σ***	**ΣΣ**

Die Summen innerhalb der Spalten ohne das Doppelsummen-Eckfeld seien – als Folge der Aufstockung – Sternchensummen. Der im Tabelleninneren gelegene obere linke Wert, sei das zu untersuchende Pivot. Es sei durch den in der selben Spalte darunter liegenden geheimen Innenwert sowie durch die beiden geheimen Summenwerte innerhalb der Zeilen quadergesichert (als Spur des vollständigen n-dimensionalen Sicherungsquaders in dieser zweidimensionalen Ebene). Eine Quadersicherung ausschließlich mit Quaderwerten im Tabelleninneren der 2-dimensionalen Tabelle scheidet hier aus, weil im Innern der Spalte neben dem Pivot zwei als Sperrpartner verbotene Tabellenwerte stehen, der Dummy- und ein offener verbotener Tabellenwert.

Zu befürchten ist nun, dass sich bei einer anderen Positionierung des Dummy-Wertes in der ersten Zeile, z.B. wie in Abb. 7.8 links neben das Pivot, alle in der Tabelle der Abbildung 7.7 als hinreichend für den Pivot-Schutz erachteten Sperreintragungen wirkungslos geworden sind. Andererseits sollten aber temporäre Umbelegungen von (n-2)-dimensionalen Tabellenfeldern mit all ihren Summenrändern in zu den Sternchensummen enthaltenden Hyper-Ebenen parallelen Hyper-Ebenen erlaubt sein, solange davon keine realen Summen betroffen sind.

Um obige Befürchtungen zu entkräften, betrachte man die Schutzfehler ε, die mit dem entsprechenden Vorzeichen der Quaderteilgesamtheit versehen, jedem Wert des zunächst undeformierten n-dimensionalen Sicherungsquaders zuzuordnen sind: Bei der n-dimensionalen Quadersicherung sei das Pivot mit dem Schutzfehler + ε behaftet. Dann kommt dem geheimen Innenwert unter dem Pivot der Schutzfehler - ε zu, damit die zugehörige Zeilensumme fehlerfrei bleiben kann. Die Schutzfehler der beiden geheimen Summenwerte sind die selben wie die der geheimen Innenwerte jeweils in der selben Zeile. ε ist der Lösungsparameter der Quadergleichungen für alle 2^n geheimen Quaderwerte als Unbekannte!

Abb. 7.8: **Zweidimensionale Spur des deformierten Sicherungsquaders**

Dummy	**Pivot + ε**	$\Sigma_{geheim} + \varepsilon$
$W_{geheim} - \varepsilon$	$Wert_{verboten}$	$\Sigma_{geheim} - \varepsilon$
$\Sigma^* - \varepsilon$	$\Sigma^* + \varepsilon$	$\Sigma\Sigma$

Bei (temporärer) Einordnung des Dummy-Werts links neben das Pivot wären die beiden Sperrungen im Tabelleninneren berechenbar; allein die Eigenschaft der Sternchensummen, nicht in der Veröffentlichungstabelle zu erscheinen, bewahrt diese Sperrungen vor direkter Rückrechenbarkeit durch Differenzbildung: Auch den betroffenen Sternchensummen Σ* ist nach dieser Umstrukturierung der Schutzfehler ε zuzuordnen (vgl. Abb. 7.8) und zwar mit jeweils dem selben Vorzeichen wie beim darüber stehenden geheimen „Quaderwert“:

Die zuletzt gewählte Anordnung der Sicherungspartner des Pivots ergibt einen geschlossenen Polygonzug von geheimen Werten in dem betrachteten 2-dimensionalen Schnitt der reduzierten Tabelle. Dabei heben sich die einander benachbarten Schutzfehler gegenseitig auf oder haben einen „Gruppenwechsel“ zwischen sich, so dass der Betrag des Schutzfehlers (innerhalb seiner durch die zu unterstellende Vorinformation zu berechnenden Schutzfehlergrenzen) beliebig frei gewählt werden kann. Ein geschlossener Polygonzug geheimer Werte alleine bietet jedoch noch keinen hinreichenden Schutz gegen Rückrechnung!

Das Pivot in der Abb. 7.7 ist als Element eines n-dimensionalen Quaders nach Voraussetzung quadergesichert. Daher gibt es noch $2^{n-2} - 1$ weitere Vierfeldertabellen mit dem selben Sperrmuster wie in Abb. 7.7 mit den selben ε-Schutzfehlern und mit den selben ε-Vorzeichen, falls die Quaderelemente in der jeweiligen Tabelle die selbe Indizierung haben (gerade oder ungerade, vgl. Definition in Abschnitt 3.1.2.2) , oder mit genau entgegengesetzten ε-Vorzeichen sonst.

Jede dieser $2^{n-2} - 1$ Tabellen entsteht aus der von Abb. 7.8 durch Umsetzen der n-2 Quaderindizes, die selbst nicht Gliederungsmerkmale der gegebene Vierfeldertabelle sind, sondern als deren Parameter fungieren, in eines der anderen $2^{n-2} - 1$ n-2-Tupel von Quaderindizes. Im diskreten Tabellenraum veranschaulicht, werden sie auf den $n - 2$ zur Ebene der Abb.7.8 senkrechten Achsen abgetragen. Die Differenz der n-2-Tupel von Parametern einer der $2^{n-2} - 1$ Tabellen und der Tabelle von Abb. 7.8 liefert den Verschiebevektor, der die Tabelle der Abb. 7.8 in die andere Tabelle überführt.

Die Gefahr, dass eine dieser Ebenen nicht alle „Gegensperrungen“ der Ebene von Abb. 7.8 enthält, besteht schon wegen der vorausgesetzten Quadersicherung nicht. Außerdem hätte eine Verschiebung dann gegenüber den Eckfeldsummen solcher Ebenen in Richtung eines der o.g. Verschiebevektoren zu erfolgen (z.B. Versatz der ersten Zeile gegen die zweite entlang des o.g. Verschiebevektors), was nur im Falle von Sternchen-Eckfeldsummen möglich gewesen wäre. In diesem Fall wären aber zwei unabhängige Sicherungsquader über die Eckfelder gegeben. Eine ähnlich einfache Situation wäre auch bei ganz im Innern der Abb.7.8-Tabelle gelegenen Quaderwerten eingetreten, wo nach der Verschiebung zwei Quader mit den Sternchensummen hätten gebildet werden können.

Weil der Pivot-Schutzfehler ε der selbe ist wie bei der Anordnung, mit der die n-dimensionale Quadersicherung vorgenommen wurde und somit der obige Tabellenschnitt hinsichtlich des Sperrmusters und seiner Schutzfehler genau in den n-dimensionalen Gesamtquader passt, ist das durch die Umsortierung erhaltene Sperrmuster für den Schutz des Pivots in der n-dimensionalen Tabelle hinreichend. Ferner treten die in Abb. 7.8 zusätzlich eingetragenen Schutzfehler von Sternchensummen sowie alle Dummy- und Sternchensummenwerte in der Veröffentlichungstabelle nicht in Erscheinung, so dass die Sperrmuster in Abb. 7.7 und Abb. 7.8 auch in Bezug auf die Veröffentlichung als völlig gleichwertig angesehen werden müssen.

D.h. (temporäre) Dummy-Verschiebungen ohne Austausch von realen Summanden zwischen den zu veröffentlichenden Tabellensummen, wenn diese (temporären) Verschiebungen also in einer zu einer Sternchensummen enthaltenden Hyper-Ebene parallelen Hyper-Ebene verlaufen und davon keine zu veröffentlichenden Tabellenwerte dieser Hyper-Ebenen betroffen sind, tun der sekundären Geheimhaltung keinen Abbruch.

7.3.5.3 Besonderheiten deformierter Quader

1. Um die für die Bewertung der Quaderelemente so wichtige Symmetrie-Eigenschaft, ob gerade oder ungerade Indizierung vorliegt, auch beim deformierten Quader beibehalten zu können, muss die Eigenschaft der geraden oder ungeraden Indizierung bei jeder Verschiebung direkt mitübertragen werden, weil sie sonst für den deformierten Quader verloren geht.

2. Die hier dargestellte Form der Quaderdeformation bezieht sich immer auf jeweils nur einen Sternchenindex. D.h. Verschiebungen von Quaderteilgesamtheiten gegenüber den restlichen Quaderelementen erfolgen immer nur innerhalb der durch diesen Sternchenindex fixierten (n-1-dimensionalen) Hyperebene von Werten der aufgestockten n-dimensionalen Tabelle. Verschoben werden dabei n-2-dimensionale Quaderteile. Diese Form der Quaderdeformation ist in dem zuletzt vom LDS NRW entwickelten EDV-Programm QUIT realisiert. Erste Auswertungsergebnisse mit simulierten Umsatzdaten sind im Anhang A 4.2 aufgeführt.

3. Prinzipiell braucht die Verschiebung von Quaderteilen aber nicht auf nur einen Sternchenindex begrenzt werden, es können auch Verschiebungen niedriger dimensionaler, durch mehr als einen Sternchenindex charakterisierte Ebenen erfolgen. Damit könnten auch kleinere als n-2-dimensionale Quaderteile verschoben werden. Diese Form der Quaderdeformation ist im zuletzt entwickelten EDV-Programm QUIT aber bisher nicht realisiert.

4. Anders als bei der Quaderauswahl ohne anschließende Deformation wird bei der Auswahl von Quadern, die anschließend noch verändert werden können, auch auf verbotene Dummies zurückgegriffen. Das bedeutet, dass verbotene Dummies bewertet werden müssen. Um aus Rechenzeitgründen möglichst wenig zu deformieren, sollten Quader mit vielen verbotenen Dummies nicht so attraktiv sein wie Quader mit wenigen oder gar keinen, ohne dabei Quader mit Randsummenwerten zu bevorzugen. Verbotene Dummies erhalten daher - ausschließlich für die Quadersummenberechnung - dimensionsabhängige positive Werte zuerkannt, die unter den Randsummenwerten liegen (vgl. Randsummengewichtung 5.2.1).

Schlussbemerkungen

1. Die natürliche Aufteilung des Geheimhaltungsproblems bei nach n Ordnungskriterien aggregierten Daten in eine Hierarchie von Unterproblemen, die zunächst unabhängig voneinander bearbeitet werden und dann durch mehrmalige Zurückführung auf die Gesamtdaten immer wieder aneinander abgeglichen werden - indem aber jedes Mal wieder das entsprechende Unterproblem für sich allein behandelt wird - liefert keine hinreichende Sicherung der Gesamtdaten. Ursache für mögliche "Geheimhaltungslücken" ist die durch Schätzfehler geheimer Randsummenwerte bedingte "Fehler-Austausch-Wechselwirkung" zwischen den Untertabellen. Diese erzwingt die Bildung gröberer Strukturen durch Zusammenfassungen von Untertabellen, das heißt Aufstockung der Dimension, so dass diese Untertabellen nur noch zu gemeinsamen Summen höherer Aggregate beitragen, die nicht mehr durch Sperrungen „aneinandergekoppelt“, d.h. voneinander abhängig gemacht sind. Die so erhaltenen vollständigen Tabellen können unabhängig voneinander gesichert werden.

   Diese Entkopplung von Untersystemen gelingt jedoch nicht im allgemeinen Falle der tabellenübergreifenden Geheimhaltung. Hier muss eine weitgehende Unterbindung von Summensperrungen in Überlappungsbereiche für eine ausreichende Entkopplung sorgen. Ein Schritt dahin ist die Einführung von Nullwerten als Sperrpartner, um so die oftmals in den Überlappungsbereich fallenden Summen-Sekundärsperrungen insbesondere der höheren Hierarchien weitgehend zu unterbinden.

2. Genau wie mehrfach durch Zwischensummen untergliederte (unvollständige) Tabellen nicht durch Aufteilung in Untertabellen mit iterativem Abgleich hinreichend gesichert werden können, verhält es sich mit mehr als zweidimensionalen Tabellen, die auch nicht durch Aufteilung in alle zweidimensionalen oder gar in alle eindimensionalen Tabellen mit iterativem Abgleich gesichert werden können. Dazu gibt es ebenfalls Gegenbeispiele. Insbesondere die Zerlegung in lauter eindimensionale Tabellen und deren gemeinsamer Abgleich führt zu Offenlegungen, weil durch diese Zerlegung und deren Abgleich genau das Differenzenverfahren realisiert wird, das bekanntlich nicht sicher ist.

   Zur Sicherung einer vollständigen Tabelle bedarf es also eines Verfahrens, das eine mehr als zweidimensionale Tabelle als Ganzes behandelt. So ein Verfahren steht mit dem vorliegenden Quaderverfahren zur Verfügung. Es vermag nicht nur n-dimensionale vollständige Tabellen hinreichend gegen eindeutige Rückrechnung seiner geheimen Werte zu sichern, sondern bietet auch einen hinreichenden Intervallschutz. Dabei ist die mathematische Struktur des Verfahrens so einfach, dass es bei kleineren Tabellen sogar manuell exakt durchgeführt werden kann, d.h. es besteht eine direkte manuelle Überprüfbarkeit.

3. Ein in letzter Zeit vermehrt diskutierter, die Wahrung der Geheimhaltung in n-dimensionalen Tabellen wesentlich verschärfender Aspekt ist die Berücksichtigung von Vorinformationen über die Tabellenwerte. Dabei handelt es sich um das Wissen, das ein Datennutzer über die Tabellendaten auch ohne deren

Kenntnis besitzt, sei es, dass ein Teil der Daten bereits in anderen Tabellen veröffentlicht worden ist, wie z.B. bei sog. überlappenden Tabellen, oder, dass der Tabellennutzer aufgrund seines Fachwissens bereits Schätzintervalle für die Tabellenwerte angeben kann.

Die gröbste Form der zuletzt genannten Vorinformation ist das Wissen, dass es sich um eine Tabelle mit nicht negativen Werten handelt, wodurch die Wahrung der Geheimhaltung bereits soweit verschärft wurde, dass nicht mehr nur die Vermeidung der eindeutigen Rückrechenbarkeit, sondern die Vermeidung der zu genauen Rückrechenbarkeit gefordert werden musste. Eine weitere Verschärfung der Sicherung sensibler Tabellendaten ergibt sich aus der Eingrenzung der Tabellenwerte durch vom Nutzer vorgebbare Schätzintervalle.

Hier tritt insofern eine ganz neue Situation auf, als es Tabellen geben kann, die bei vorgegebenen relativen Mindestspannweiten zum Schutze primär geheimer Werte gar nicht mehr gesichert werden können, wenn etwa das vom Nutzer angebbare Schätzintervall eine kleinere Spannweite besitzt als das mit der relativen Mindestspannweite für den Schutz vorgegebene Intervall für die Quaderauswahl. Eine Sicherung zu genau vorbestimmter Tabellenwerte ist dann aber auch mit keinem anderen Verfahren zur sekundären Geheimhaltung möglich!

Aus diesem Grunde werden sich auch die Sicherungsmöglichkeiten von Statistiktabellen, deren Werte sich u.U. durch Schätzfehler mit anderen bereits veröffentlichten Tabellen stark einengen lassen, wie z.B. bei kurzzeitig aufeinanderfolgenden Zeitreihentabellen, sehr in Grenzen halten. Eine i.A. wohl praktikablere Lösung bietet da die im nächsten Punkt angesprochene Gewichtung von Tabellenwerten.

Darüber hinaus ist anzumerken, dass bei vorausgesetzter Vorinformation in Gestalt von Schätzintervallen auch Tabellen mit nicht ausschließlich positiven Werten und Nullen mit Intervallschutz gesichert werden müssen: Wurden Tabellen mit positiven und negativen Werten bisher so behandelt, als fehlte die Information über eine mögliche Eingrenzung der Werte durch den Tabellennutzer in Form der Positivität der Tabelle, so dass die Verhinderung der eindeutigen Rückrechenbarkeit genügt hätte, so muss bei Vorliegen von Schätzintervallen auch bei Tabellen mit positiven und negativen Werten die Quaderauswahl mit range-Kriterium durchgeführt werden.

4. Eine ganz wesentliche Erweiterung des Anwendungsspektrums des Quaderverfahrens wurde durch die Einführung der externen Gewichtung der Tabellenwerte erreicht. Der durch die Sperrungen von Tabellenwerten verursachte erfassbare Informationsverlust muss nicht mehr wie bisher allein durch den Betrag des Wertes bestimmt sein, er lässt sich nun durch Vorgabe von Gewichten in weiten Grenzen modifizieren, ohne dabei auf den üblichen tabellenwertebezogenen Intervallschutz verzichten zu müssen. Bei umfangreichen Tabellen wird man die für jeden Tabellenwert einzeln vorgebbaren Gewichte in praktikabler Weise als Funktionen fachbezogener Variabler eintragen. Damit können beispielsweise fallzahlabhängige, tabellenfeldabhängige, aber auch von externen Angaben abhängige Gewichtungen vorgenommen werden.

   In diesem Zusammenhang ist als besonders praxisrelevantes Beispiel die Bearbeitung von kurzzeitig aufeinanderfolgenden Zeitreihentabellen zu nennen. Obwohl der Parameter Zeit im Sinne der sekundären

Geheimhaltung keine zusätzliche Tabellendimension darstellt, denn es wird nicht über die Zeit summiert, erlauben gerade „Längsschnittauswertungen“ – z.B. durch Schätzen von „Antwortausfällen“ aus Vorperiodenwerten – u.U. zu genaue Rückschlüsse auf geheime Tabellendaten. Abhilfe schafft nun die sekundäre Geheimhaltung mit einer den Schätzfehler von Längsschnittauswertungen berücksichtigenden extern vorgebbaren Gewichtsfunktion, die das Sperrmuster so justiert, dass auch aus Vorperiodenwerten zu genau zu berechnende sensible Werte noch ausreichend gesichert werden (z.B. durch fortlaufende Sperrungen einander entsprechender Werte über mehrere Zeitperioden).

5. Mit vollständigen Quadern gesicherte geheime Tabellenwerte können nicht genauer bestimmt werden, als es die für die Quaderauswahl vorgegebene relative Mindestspannweite erlaubt (vgl. 3.2.3.3). Das gilt auch unter Berücksichtigung von Vorwissen. Damit ist die Möglichkeit gegeben, die Schutzintervalle durch vollständige Quader gesicherter Werte zu veröffentlichen, ohne die Geheimhaltung dieser Werte zu gefährden. Bei Quaderüberlappungen werden die vereinigten Schutzintervalle der überlappenden Quaderwerte ausgegeben.

   Die nachfolgende mit dem EDV-Programm QUIT bearbeitete Beispieltabelle dokumentiert den Nutzer-Informationsgewinn durch die Ausgabe von Schutzintervallen sehr eindrucksvoll. Besonders durch die beiden Schutzintervalle in der Zwischensummen-Zeile 130 werden Angaben von Sekundärsperrungen weitgehend offengelegt, ohne den Schutz der Primärsperrungen zu gefährden: Die sekundär geheimen Werte 3031 und 1672 werden bis auf relative Schutzintervalllängen von 4 % bzw. 7 %, d.h. bis auf relative Fehler von ± 2 % bzw. ± 4 % genau angeben, obwohl die Primärsperrungen durch eine relative Mindestspannweite von 50% geschützt sind .

   Durch vollständige Quader gesicherte geheime Tabellenwerte brauchen also in ihrer Veröffentlichungstabelle nicht mehr durch Schutzsternchen dargestellt, sie können statt dessen durch ihre Schutzintervalle repräsentiert werden, was der angestrebten Akzeptanzförderung ganz sicher dienlich sein wird.

Beispieltabelle mit Ausgabe von Schutzintervallen (positive Tabelle, 50 % relative Mindestspannweite)

1. Schlüssel	2. Schlüssel: ACD	ACC	ACB	ACA	AC	ABC	ABB	ABA	AB	AAD	AAC	AAB	AAA	AA	A
134	0 - 122 5 S	0 - 122 2 P	1.445 20	549 12	2.116 39	4.128 34	345 15	211 12	4.684 61	65 - 355 21 S	0 0	0 0	61 - 351 2 P	416 23	7.216 123
133	30 - 106 1 P	0 - 76 4 S	0 0	23 3	129 8	2.567 44	2.332 30	432 21	5.331 95	732 51	644 34	0 0	0 0	1.376 85	6.836 188
132	723 9	254 11	327 5	543 19	1.847 44	1.123 64	4.427 59	1.632 26	7.182 149	432 23	0 0	234 36	0 0	666 59	9.695 252
131	2.156 33	1.342 23	1.111 17	99 4	4.708 77	590 11	2.334 28	342 9	3.266 48	0 - 290 3 S	0 0	0 0	0 - 290 17 S	290 20	8.264 145
130	2922 - 3040 48 S	1664 - 1782 40 S	2.883 42	1.214 38	8.800 168	8.408 153	9.438 132	2.617 68	20.463 353	1.519 98	644 34	234 36	351 19	2.748 187	32.011 708
125	302 - 332 5 S	0 - 30 3 S	411 18	0 0	743 26	0 0	56 5	0 0	56 5	712 50	3.421 84	0 0	0 0	4.133 134	4.932 165
124	37 - 68 4 S	0 - 31 1 P	2.152 29	399 11	2.619 45	0 0	123 10	0 0	123 10	345 44	2.612 61	55 3	0 0	3.012 108	5.754 163
123	99 8	311 10	754 19	345 16	1.509 53	221 7	0 - 107 2 P	0 - 107 6 S	328 15	123 23	321 41	567 32	43 4	1.054 100	2.891 168
122	1826 - 1855 33 S	0 - 31 1 P	88 4	0 0	1.944 38	0 0	621 13	0 0	621 13	1.015 89	2.221 52	96 18	641 8	3.973 167	6.538 218
121	344 15	298 13	0 0	934 9	1.576 37	0 0	74 8	0 0	74 8	0 0	231 33	0 0	875 72	1.106 105	2.756 150
120	2.657 65	651 28	3.405 70	1.678 36	8.391 199	221 7	874 - 981 38 S	0 - 107 6 S	1.202 51	2.195 206	8.806 271	718 53	1.559 84	13.278 614	22.871 864
113	43 - 165 2 P	109 - 231 8 S	29 3	1.001 19	1.304 32	0 0	0 0	0 0	0 0	0 - 32 2 P	0 0	0 - 32 2 P	0 0	32 4	1.336 36
112	423 18	0 0	0 0	0 0	423 18	0 0	188 - 295 5 S	0 - 107 2 P	295 7	745 71	0 0	67 8	0 0	812 79	1.530 104
111	28 5	0 0	0 0	0 0	28 5	0 0	0 0	0 0	0 0	127 - 159 25 S	0 0	70 - 102 7 S	0 0	229 32	257 37
110	494 - 616 25 S	109 - 231 8 S	29 3	1.001 19	1.755 55	0 0	188 - 295 5 S	0 - 107 2 P	295 7	904 98	0 0	169 17	0 0	1.073 115	3.123 177
100	6.192 138	2.544 76	6.317 115	3.893 93	18.946 422	8.629 160	10.607 175	2.724 76	21.960 411	4.618 402	9.450 305	1.121 106	1.910 103	17.099 916	58.005 1.749

Legende:

	1. Aggr.	2. Aggr.	3. Aggr.
Wert	10.000	10.000	10.000
Berichtspfl.	100	100	100

Berichtspfl. 0 - 122 / 100 P — untere und obere Schätzintervallgrenze; Sperrvermerk (P=primär, S=sekundär)

Anhang

A Anwendung des Quaderverfahrens auf Realdaten

A.1 Umsatzsteuerstatistik NRW 1994[10], eine umfangreiche Tabelle

Eine in Bezug auf die Gliederungsstruktur der mit dem Geheimhaltungsverfahren zu bearbeitenden Tabellen repräsentative Statistik ist der „steuerbare Umsatz“. Es handelt sich dabei um eine zweidimensionale nach regionaler Gliederung und nach wirtschaftlicher Systematik gegliederte Tabelle mit nicht negativen Werten. Die wichtigsten strukturellen Daten dieser Statistik sind in folgender Übersicht zusammengestellt:

Ausgangsdaten:

Datensätze (Tabellenfelder)	717 914
primär geheime Werte	159 051
leere Tabellenfelder	457 258
Aggregations-Niveaus in regionaler Gliederung	4
Aggregations-Niveaus in wirtschaftlicher Gliederung	7
Untertabellen	30 488

[10] Diese Auswertungen wurden auf IBM-9672 unter OS 390 mit der EDV-Programmversion GHQUAR.3 gemacht und zusammen mit den Auswertungen der Fremdenverkehrsstatistik vom Verfasser bereits 1999 im Forum der Bundesstatistik, Band 31/1999 veröffentlicht. Der Untertabellenabgleich erfolgte ohne Schutzintervallübertrag (vgl. 3.2.3).

Informationen zur Sekundärsperrung

Rechenzeit (CPU-Zeit) bei	5min30
relativer Mindestspannweite gleich 0 (ohne Intervallschutz)	
Rechenzeit (CPU-Zeit) bei	5min52
relativer Mindestspannweite gleich 1,5	
sekundäre Sperrungen bei	30 294
- relativer Mindestspannweite gleich 0	
- gesetzter Randschranke für beide Dimensionen.[11]	
sekundäre Sperrungen bei	51 351
- relativer Mindestspannweite gleich 1,5	
- gesetzter Randschranke für beide Dimensionen.	

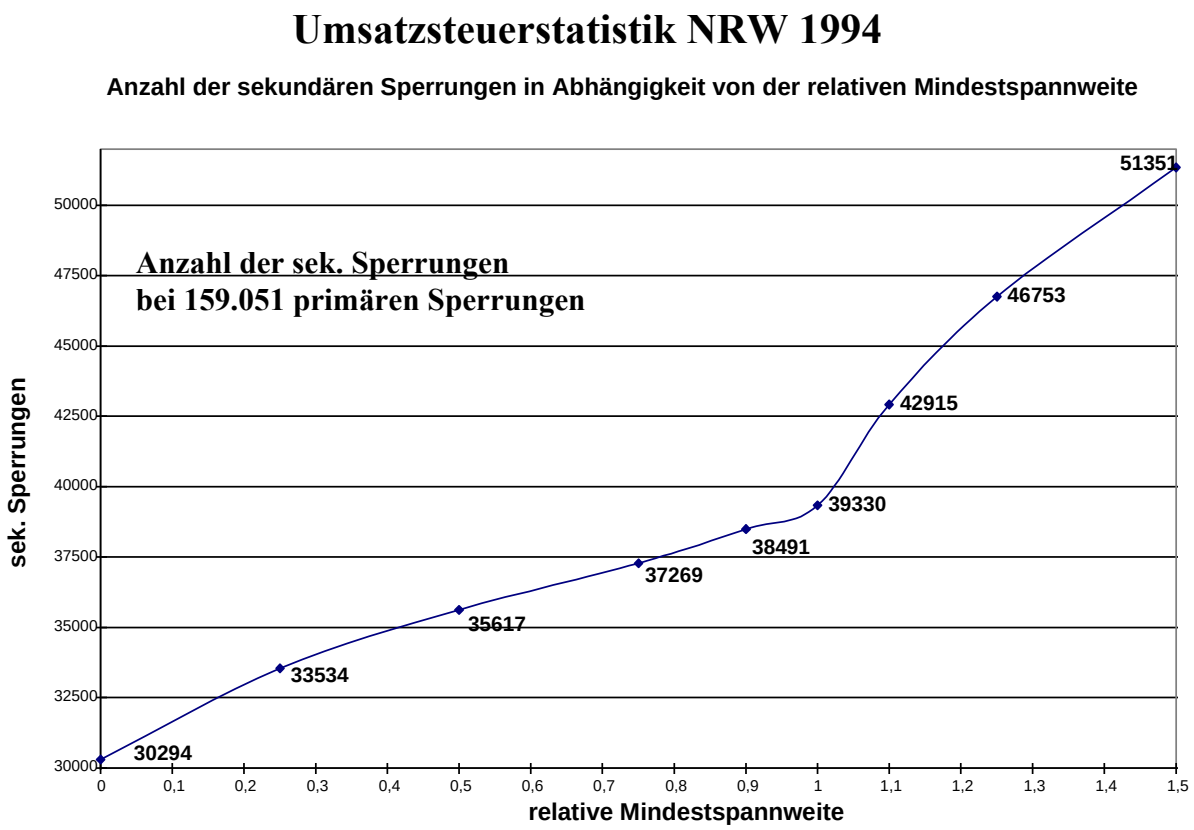
Die Umsatzsteuerstatistik ist mit ihren mehr als 700 000 Datensätzen eine im Vergleich zu anderen Statistiken des Landes Nordrhein-Westfalen sehr umfangreiche Tabelle, die besonders fein gegliedert ist. Die sehr feine Gliederung äußert sich in der sehr schwachen Besetzung mit weit über die Hälfte leeren Tabellenfeldern und in der sehr hohen Anzahl von Primärsperrungen, die mehr als die Hälfte der besetzten Tabellenfelder ausmachen. Die Feinheit der Gliederung äußert sich aber auch in der großen Anzahl von mehr als 30 000 Untertabellen, die alle aneinander abgeglichen werden müssen.

Die obige Übersicht umfasst zwei unabhängig voneinander durchgeführte Sperrvorgänge: Beim ersten Sicherungslauf wurde die relative Mindestspannweite zur Auswahl von Sicherungsquadern gleich Null vorgegeben, beim zweiten gleich 1,5 gesetzt. Die unterschiedliche Wahl der relativen Mindestspannweite hat folgende Auswirkungen, die sich in obiger Übersicht niederschlagen: Während bei einer relativen Mindestspannweite von Null alle Quader mit Nullwerten in nur einer Quaderteilgesamtheit zur Sicherung primär geheimer Werte in Betracht kommen, dürfen bei Vorgabe einer von Null verschiedenen Spannweite nur solche Quader zur Sicherung herangezogen werden, deren Spannweite bezogen auf den zu schützenden geheimen Wert mindestens so groß wie die vorgegebene relative Mindestspannweite ist. Das hat zur Folge, dass bei von Null verschiedener Mindestspannweite häufiger auf Randsummenwerte der betreffenden Untertabelle ausgewichen werden muss als bei Quaderauswahl ohne

[11] Die für jede Dimension eingeführte Randschranke dient der Justierung, insbesondere bei überlappenden Tabellen: Bei erforderlichen Randsummensperrungen werden Summen mit Randschranke weitgehend gemieden.

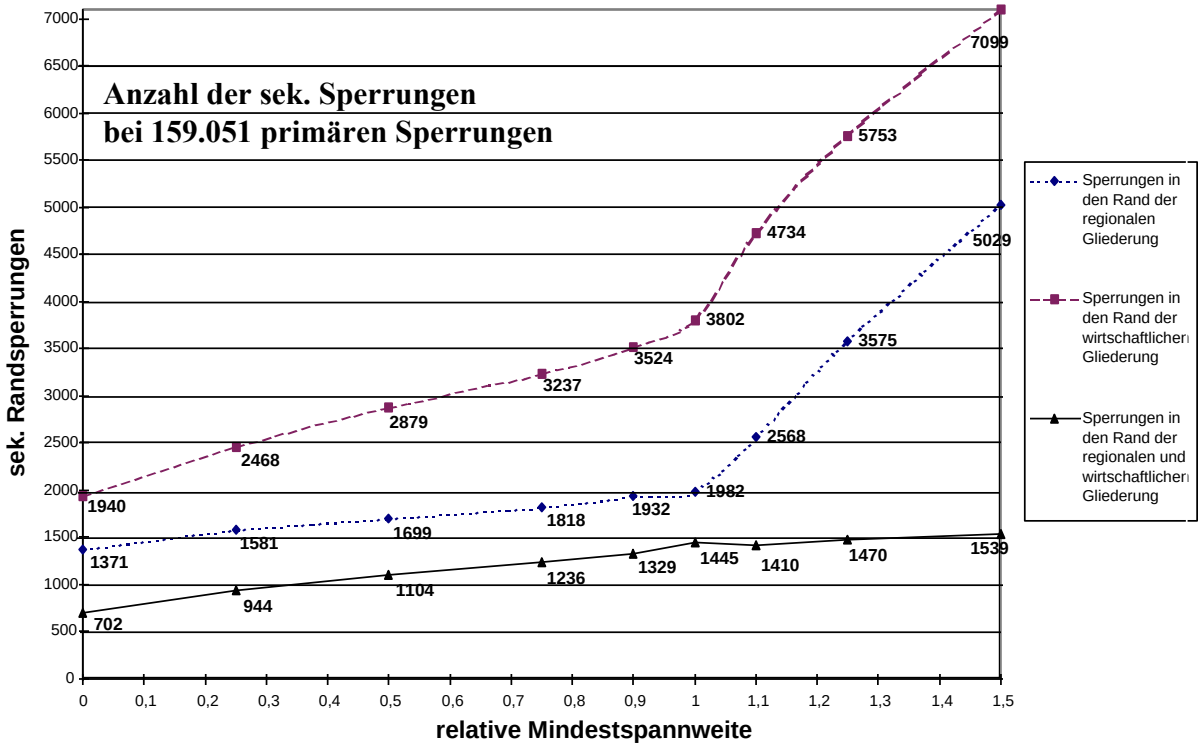
Berücksichtigung einer Mindestspannweite. Dadurch werden beim Sichern mit relativer Spannweite 1,5 mehr Sperrungen (und auch eine höhere Summe zu sperrender Werte) erzwungen als bei fehlender Spannweitenvorgabe, weil Summensperrungen in den zugehörigen Untertabellen höherer Hierarchiestufe gesichert werden müssen.

Dieses Verhalten bestätigt sich in den beiden nachstehenden graphischen Darstellungen:



Umsatzsteuerstatistik NRW 1994

Anzahl der sek. Randsperrungen in Abhängigkeit von der relativen Mindestspannweite



Die monoton mit der relativen Mindestspannweite steigende Zahl von Sekundärsperrungen insgesamt (erste Umsatzsteuerstatistik-Darstellung) und in den Randsummen der Untertabellen (zweite Darstellung) weist bei einer relativen Mindestspannweite von 1 einen Verlaufsknick auf (mit Ausnahme der Eckfeldsperrungen), der eine deutliche Zunahme von Sperrungen und insbesondere von Randsperrungen bei über 1 hinausgehenden zunehmenden relativen Spannweiten anzeigt. Dies ist darauf zurückzuführen, dass oberhalb von 1 mehr Untertabellen vorhanden sind, bei denen nur noch durch Quader mit zwei Randwerten die Quaderauswahl-Bedingung erfüllt werden kann. Das hat insbesondere im Bereich der höher aggregierten Tabellen viele Sekundärsperrungen durch Untertabellenabgleich zur Folge. Dem gemäß beobachtet man bei Gliederungen mit nur 2 Aggregationsstufen keinen so ausgeprägten Knick bei 100 % relative Mindestspannweite; die Fremdenverkehrsstatistik (Anhang A.2) ist dafür ein Beispiel.

Dennoch bleibt die Gesamtzahl von Sekundärsperrungen auch im Falle der großen relativen Mindestspannweite von 150 % deutlich hinter den Primärsperrungen zurück: So kommen im Durchschnitt im hier ungünstigsten Fall ca. 5 Primärsperrungen aber nur etwa eine Sekundärsperrung auf jede der 30 000 Untertabellen. Dies macht deutlich, dass sich selbst gröbere Abweichungen von der Optimalität des Sicherungsverfahrens in der Gesamtzahl der Sperrungen nur marginal bemerkbar machen - was wiederum für den Einsatz eines heuristischen Sperrverfahrens wie des Quaderverfahrens spricht.

A.2 Fremdenverkehrsstatistik NRW 1995[12], überlappende Tabellen

Als zweite Anwendung wurde die Fremdenverkehrsstatistik gewählt, weil sie aus drei einzelnen, aber einander überlappenden dreidimensionalen Tabellen besteht. Die Statistik ist daher beispielhaft für die Geheimhaltung in mehr als zweidimensionalen Tabellen mit nicht negativen Werten, die außerdem noch mit anderen Tabellen gewisse Tabellenfelder gemeinsam haben.

Die wichtigsten Parameter sind, wie beim ersten Beispiel, in Form von Übersichten zusammengestellt, und zwar für die gesamte, aus allen drei Tabellen bestehende Statistik und für jede Tabelle einzeln. Der Gesamttabellenübersicht folgt eine graphische Darstellung der Anzahl der Sekundärsperrungen in Abhängigkeit von der relativen Mindestspannweite.

Gesamttabelle:

Ausgangsdaten:

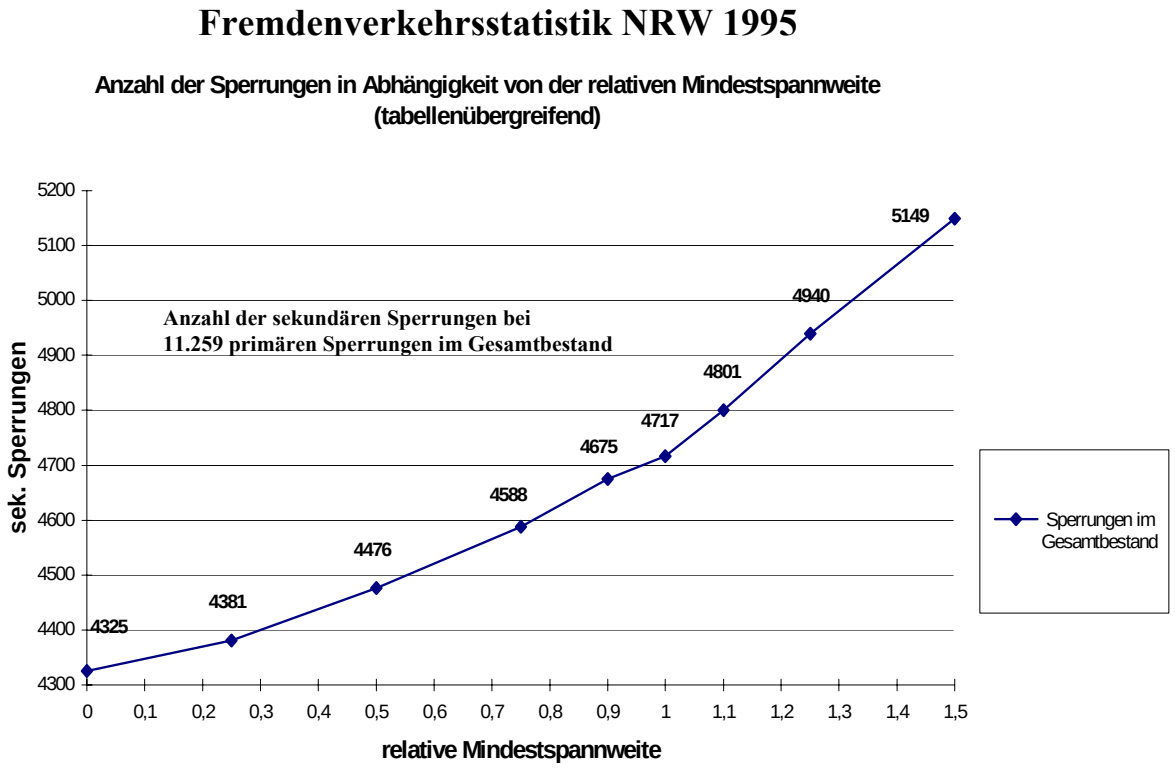
Datensätze (Tabellenfelder)	77 940
primär geheime Werte	11 259
leere Tabellenfelder	45 152

Informationen zur Sekundärsperrung

sekundäre Sperrungen bei	4 325
- relativer Mindestspannweite gleich 0 (ohne Intervallschutz)	
- gesetzter Randschranke für die 1. und 2. Dimension	
sekundäre Sperrungen bei	5 149
- relativer Mindestspannweite gleich 1,5	
- gesetzter Randschranke für die 1. und 2. Dimension	

[12] Die Auswertungen wurden mit dem im LDS NRW entwickelten EDV-Programm GHMITER.21 gemacht, das auf das durch GHQUAR.3 realisierte Quaderverfahren zurückgreift und das mit verschränkten Einzeltabellen arbeitet (siehe 6.1.1). Der Tabellen- und Untertabellenabgleich erfolgte ohne Schutzintervallübertrag (vgl. 3.2.3).

Anzahl der Tabellen mit je 3 Dimensionen	3
Rechenzeit (CPU-Zeit), einheitlich für alle Veränderungen der Mindestspannweite	3min30



An die drei Einzeltabellen-Übersichten schließt sich ein gemeinsames Schaubild an, in dem die Anzahl der Sekundärsperrungen als Funktion der relativen Mindestspannweite für jede der drei Tabellen einzeln ausgewiesen wird.

1. <u>Einzeltabelle</u>

Ausgangsdaten:

Datensätze (Tabellenfelder)	23 382
primär geheime Werte	3 465

leere Tabellenfelder	17 164
Aggregations-Niveaus in regionaler Gliederung	4
Aggregations-Niveaus der Betriebsart	2
Aggregations-Niveaus der Ausstattungsklasse	2
Untertabellen	37

Informationen zur Sekundärsperrung

sekundäre Sperrungen bei - relativer Mindestspannweite gleich 0 (ohne Intervallschutz) - gesetzter Randschranke für die 1. und 2. Dimension	2 104
sekundäre Sperrungen bei - relativer Mindestspannweite gleich 1,5 - gesetzter Randschranke für die 1. und 2. Dimension	2 316

2. <u>Einzeltabelle</u>

Ausgangsdaten:

Datensätze (Tabellenfelder)	31 176
primär geheime Werte	4 826
leere Tabellenfelder	23 247
Aggregations-Niveaus in regionaler Gliederung	4
Aggregations-Niveaus der Betriebsart	2
Aggregations-Niveaus der Bettenbestandsgrößenklassen	2
Untertabellen	37

Informationen zur Sekundärsperrung

sekundäre Sperrungen bei	2 211
- relativer Mindestspannweite gleich 0 (ohne Intervallschutz)	
- gesetzter Randschranke für die 1. und 2. Dimension	
sekundäre Sperrungen bei	2 540
- relativer Mindestspannweite gleich 1,5	
- gesetzter Randschranke für die 1. und 2. Dimension	

3. Einzeltabelle

Ausgangsdaten:

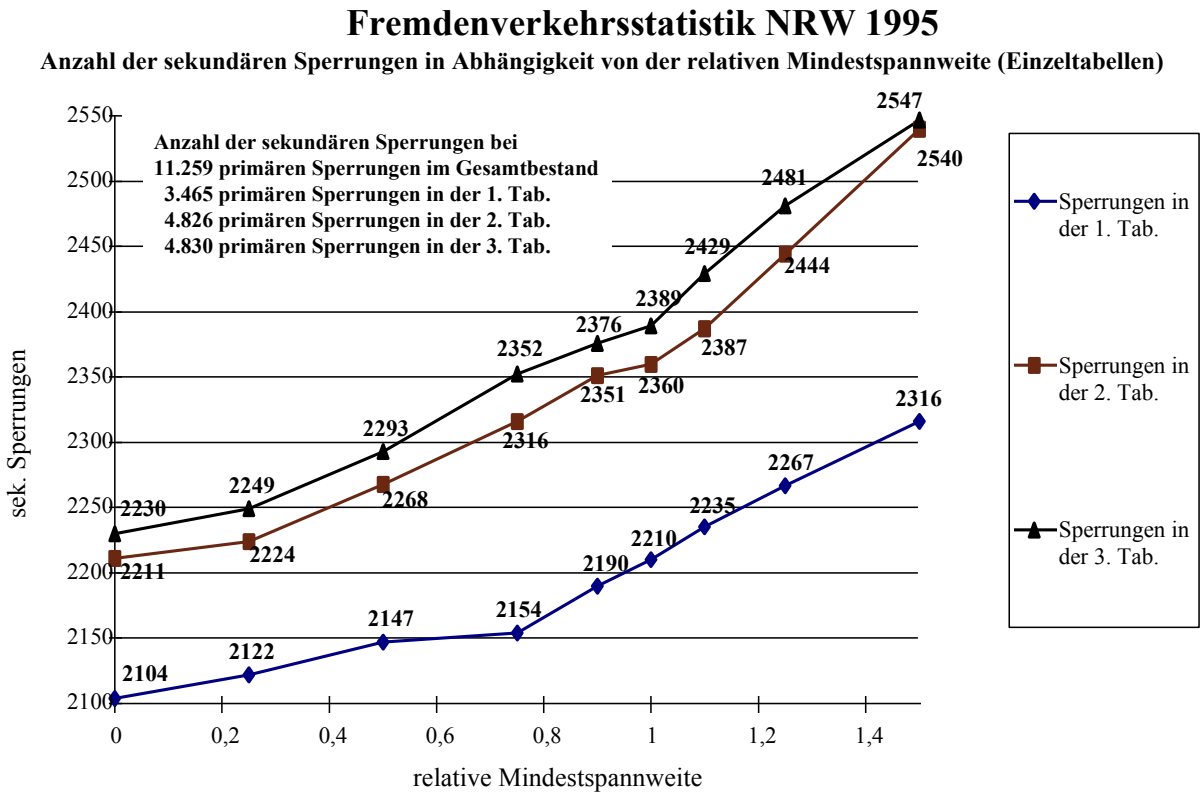
Datensätze (Tabellenfelder)	31 176
primär geheime Werte	4 830
leere Tabellenfelder	23 235
Aggregations-Niveaus in regionaler Gliederung	4
Aggregations-Niveaus der Betriebsart	2
Aggregations-Niveaus der Bettenangebotsgrößenklassen	2
Untertabellen	37

Informationen zur Sekundärsperrung

sekundäre Sperrungen bei	2 230
- relativer Mindestspannweite gleich 0 (ohne Intervallschutz)	
- gesetzter Randschranke für die 1. und 2. Dimension	
sekundäre Sperrungen bei	2 547

- relativer Mindestspannweite gleich 1,5

- gesetzter Randschranke für die 1. und 2. Dimension



Wie aus den Strukturdaten der Übersichten zu entnehmen ist, handelt es sich um drei völlig gleich strukturierte Tabellen, die sich lediglich in ihrem dritten Gliederungskriterium voneinander unterscheiden. Dem gemäß ergeben sich auch für die Anzahl der Sekundärsperrungen in Abhängigkeit von der relativen Mindestspannweite in der letzten Darstellung drei gleichartige monotone Kurvenverläufe, die sich auch in Bezug auf die Anzahl der Sekundärsperrungen nur geringfügig voneinander unterscheiden.

Die zur Gesamttabelle gehörige Darstellung zeigt einen im Vergleich zu den Einzelkurven mittleren Verlauf, wobei die einzelnen in die Darstellung eingetragenen Sekundärsperrungen nicht durch Addition der Sekundärsperrungen der Einzelkurven berechnet werden können, weil einige Tabellenfelder allen Tabellen gemeinsam angehören und daher in der Gesamtübersicht und - Darstellung auch nur einmal aufgeführt werden.

Als bemerkenswert erscheint an diesen Tabellen die im Vergleich zu anderen Tabellen der amtlichen Statistik verhältnismäßig große Anzahl von Sekundärsperrungen, die bei allen Tabellen schon etwa halb so groß wie die Anzahl der Primärsperrungen ist. Dass die Anzahl von Sekundärsperrungen in einer dreidimensionalen Tabelle vergleichsweise höher ausfallen muss, als in einer zweidimensionalen mit sonst vergleichbarer Tabellenbesetzung, ergibt sich aus der höheren Anzahl von Quaderwerten: Im dreidimensionalen Fall werden 7 gesperrte Tabellenfel-

der zum Schutze eines geheimen Wertes benötigt, im zweidimensionalen Fall sind es nur drei, so dass auch bei sich bereits gegenseitig schützenden Primärsperrungen tendenziell in dreidimensionalen Tabellen immer noch ca. doppelt so viele Sekundärsperrungen hinzunehmen sein werden wie in zweidimensionalen.

Um Sekundärsperrungen in die Überlappungsbereiche der Einzeltabellen nach Möglichkeit zu verhindern, werden durch Gewichtung der Randsummen jeder Untertabelle mit Hilfe der so genannten „Randschranken“ die Summenwerte so weit erhöht, dass sie das Quaderauswahlverfahren weitgehend meidet und statt dessen auf andere nicht mit Randschranken belegte Summen ausweicht. Mit diesem Vorgehen wird von der bereits in der Einführung angesprochenen und in Abschnitt 5.2.1 näher erläuterten Modifikationsmöglichkeit der Eingabedaten zum Zwecke einer Justierung der Verteilung von Sekundärsperrungen Gebrauch gemacht.

In der vorliegenden Fremdenverkehrsstatistik wären demnach alle durch Aufsummieren der Tabellenwerte über das jeweils dritte Gliederungskriterium (in der dritten Dimension) gebildeten Randsummen mit einer positiven Randschranke zu versehen und die Randsummen bezüglich der beiden ersten Gliederungskriterien unbehelligt zu lassen.

Die Vermeidung von Sekundärsperrungen in die Überlappungsbereiche - hier die zweidimensionale, nur nach den ersten beiden Gliederungskriterien gegliederte Tabelle - garantiert jedoch keine besonders kleine Anzahl von Sekundärsperrungen in der Gesamtstatistik, wie das vorliegende Beispiel zeigt! Belegt man nur das jeweils dritte Gliederungskriterium mit einer Randschranke, wie es der Schutz der Überlappungstabelle gegen Sekundärsperrungen erfordert, und gibt man die anderen beiden Randsummen für Sperrungen frei, erhält man bei einer Mindestspannweite 0 zwar nur 285 Sekundärsperrungen in den Überlappungsbereich, muss aber 4 952 sekundäre Sperrungen in der Gesamtstatistik hinnehmen, während bei Belegung der ersten beiden Dimensionen mit Randschranken 317 Sperrungen in die Überlappungstabelle vorgenommen werden, dafür werden aber insgesamt nur 4 325 Sekundärsperrungen in die Fremdenverkehrsstatistik eingetragen, 627 weniger als beim „Schutz“ des Überlappungsbereichs.

Dass die Freigabe der Tabellensummen für Sekundärsperrungen in der regionalen Gliederung zu einer wesentlichen Erhöhung der sekundären Sperreintragungen insgesamt führt, liegt in der Feinheit der Regionalstruktur begründet, die sich über vier Aggregationsstufen erstreckt. Die beiden anderen Gliederungen der Einzeltabellen weisen dagegen nur 2 Aggregationsniveaus auf. Trotzdem ist auch die Gliederung nach Betriebsart, die 2. Dimension, mit einer Randschranke zu versehen: Die nach Auswertung der Fremdenverkehrsstatistik mit allen denkbaren Randschranken-Belegungen erhaltene hinsichtlich der Gesamtzahl sekundärer Sperrungen günstigste Randbelegung ist die in den Übersichten und graphischen Darstellungen angegebene.

A.3 Berücksichtigung von externen Schätzintervallen

Für eine empirische Untersuchung der Auswirkungen von Vorinformationen in Gestalt von externen Schätzintervallen auf die sekundäre Geheimhaltung bietet sich wieder der „steuerbare Umsatz“ NRW 1994 an, weil davon auszugehen ist, dass gerade diese sensiblen Daten den Tabellennutzern von miteinander konkurrierenden Unternehmen zumindest bis auf Schätzintervalle genau bekannt sind. Um zu zeigen, wie stark die Vorgabe externer Schätzintervalle die Sekundärsperrungen beeinflusst, wurden Sicherungsläufe zu vorgegebenen Schätzfehlern von 50 %, 100 %, 200 % und 400 % durchgeführt, und die Ergebnisse zusammen mit den Sekundärsperrungen des steuerbaren Umsatzes als positive Tabelle ohne Schätzintervalle zum Vergleich grafisch dargestellt. Beim Untertabellenabgleich wurden keine Schutzintervalle übertragen (vgl. 3.2.3).

Bei 100% überschreitenden Fehlergrenzen liegt die untere Schätzintervallgrenze im Bereich negativer Werte. In positiven Tabellen wie der des steuerbaren Umsatzes haben daher Fehlerangaben über 100% nur dann einen Sinn, wenn sie sich auf die obere Schätzintervallgrenze beziehen, die untere ist dann immer als Null anzunehmen. Die Beziehungen (10) und (11) des Abschnitts 3.2.2.2, nach denen hier die Quaderspannweite berechnet wurde, berücksichtigen diese Asymmetrie, indem sie das Schutzintervall als Schnittmenge aus dem symmetrischen, mit (9), Abschnitt 3.2.2.1 zu berechnenden Intervall und dem asymmetrischen Schutzintervall einer positiven Tabelle (gemäß (4), Abschnitt 3.1) bestimmen[13]. Die über die Vorinformation, dass es sich um eine positive Tabelle handelt, hinausgehende Einengung der Tabellenwerte durch Schätzfehlerangaben über 100% betreffen also nur noch die obere Schätzintervallgrenze und die kann beliebig hoch sein.

13

Für relative Schätzfehler $f \geq 1$ gilt nach (7a), (7b)

$$\varepsilon_{+} = \min_{X \in Q_g} [f \min X, \min_{X' \in Q_u} X'] \geq \min_{X' \in Q_u} [\varepsilon_Q, \min X'] = \varepsilon_{Q+}$$

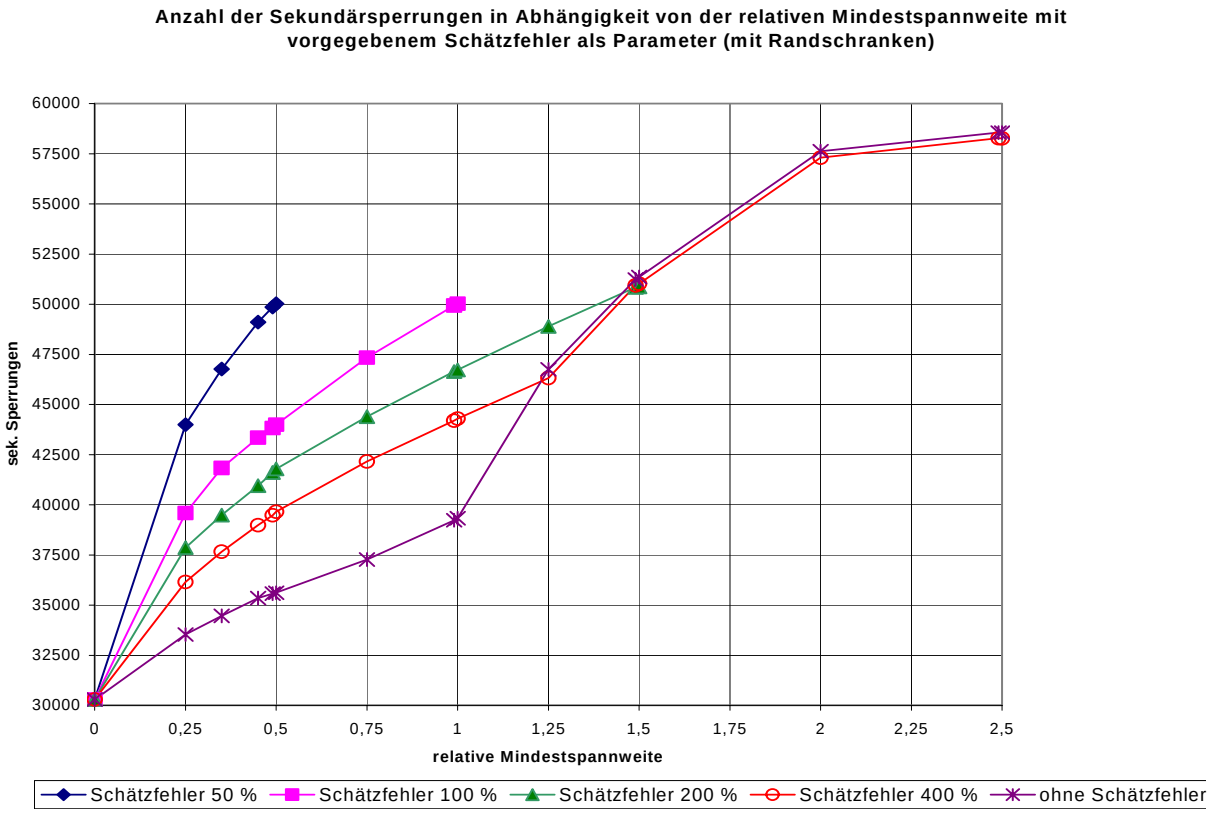
$$\varepsilon_{-} = \min_{X' \in Q_u} [f \min X', \min_{X \in Q_g} X] \geq \min_{X \in Q_g} [\varepsilon_Q, \min X] = \varepsilon_{Q-},$$

wenn hier $\varepsilon_Q = f \min_{X \in Q_g \cup Q_u} X$ gesetzt wird

Für $f < 1$ gilt nach (7a), (7b) entsprechend

$$\varepsilon_{+} = \min_{X \in Q_g} [f \min X, \min_{X' \in Q_u} f \min X'] = f \min_{X \in Q_g} [\min X, \min_{X' \in Q_u} X'] = f \min_{X \in Q_g \cup Q_u} X = \varepsilon_Q$$

analog $\varepsilon_{-} = f \min_{X \in Q_g \cup Q_u} X = \varepsilon_Q$ wie es bereits mit (9) dargestellt wurde.



Die Abbildung zeigt die Sekundärsperrungen in Abhängigkeit von der relativen Mindestspannweite mit vorgegebenem relativen Schätzfehler als Parameter. Man sieht, dass sich der Kurvenverlauf mit zunehmendem relativen Schätzfehler immer mehr abflacht, bis die Kurven bei ganz großen Schätzintervallen mit den Werten der Kurve der positiven Tabelle, bei der man Schätzfehler als beliebig groß annehmen kann, annähernd zusammenfallen, wenn sie nicht schon vorher abbrechen.

Die oben genannte Eigenschaft der Anzahlen von Sekundärsperrungen, bei kleineren Schätzintervallen mit der relativen Mindestspannweite schneller zuzunehmen als bei größeren, trifft genau die Erwartung, weil bei stärker eingeengten Tabellenwerten weniger Werte zur Sicherung eines primär geheimen Wertes zur Auswahl stehen, und somit bei größer werdender relativer Mindestspannweite öfter auf Randsummen ausgewichen werden muss - mit allen daraus resultierenden Konsequenzen (siehe dazu A.1). Dem gemäß nähern sich die Kurven mit zunehmendem Schätzintervallparameter auch immer mehr der Kurve der positiven Tabelle ohne Vorgabe von Schätzintervallen, wobei auch die Stelle des charakteristischen Verlaufsknicks sich auf eine relative Mindestspannweite von 1 zu bewegt, bis beide Kurven vollständig zusammen fallen (das wurde für einen Schätzfehler von 100.000 % tatsächlich gemessen).

Überraschend ist aber, dass es bei großen Schätzintervallparametern Kurvenabschnitte geben kann - z.B. für einen Schätzfehlerparameter von 400 % ab einer relativen Mindestspannweite von 1,25 -, bei denen die Anzahl von Sekundärsperrungen der durch das externe Schätzintervall eingeengten sekundären Geheimhaltung kleiner sind als bei der nur positiven Tabelle.

Diese Eigenart hängt mit dem Untertabellenabgleich zusammen und damit, dass Übersperrungen, die in vorangegangenen Iterationsschritten entstanden sind, nicht wieder rückgängig gemacht werden. So kann es unter Umständen günstiger sein, wenn bereits im ersten Iterationsschritt mehr Sperrungen in den Rand erfolgen, damit bei den darauf folgenden Schritten um so weniger Rücksperrungen vom Rand ins Tabelleninnere notwendig sind, die in der Regel zu Übersperrungen im Inneren der Tabelle führen. Diese Vorgänge sind äußerst komplex; sie werden daher anhand der schon mehrfach verwendeten sehr kleinen, mehrfach durch Zwischensummen unterteilten Beispieltabelle verdeutlicht.

		2. Schlüssel														
		ACD	ACC	ACB	ACA	AC	ABC	ABB	ABA	AB	AAD	AAC	AAB	AAA	AA	A
1. Schlüssel	00000134	112 5 S	10 2 P	1.445 20	549 12	2.116 39 S	4.128 34	345 15	211 12	4.684 61	321 21 S2	0 0	0 0	95 2 P	416 23 S	7.216 123
	00000133	40 1 P	66 4 S	0 0	23 3	129 8 S	2.567 44	2.332 30	432 21	5.331 95	732 51 S	644 34	0 0	0 0	1.376 85 S	6.836 188
	00000132	723 9	254 11	327 5	543 19	1.847 44	1.123 64	4.427 59	1.632 26	7.182 149	432 23	0 0	234 36	0 0	666 59	9.695 252
	00000131	2.156 33	1.342 23 S	1.111 17	99 4	4.708 77 S	590 11	2.334 28	342 9	3.266 48	34 3 S	0 0	0 0	256 17 S	290 20 S	8.264 145
	00000130	3.031 48	1.672 40	2.883 42	1.214 38	8.800 168	8.408 153	9.438 132	2.617 68	20.463 353	1.519 98	644 34	234 36	351 19	2.748 187	32.011 708
	00000125	321 5	11 3	411 18	0 0	743 26	0 0	56 5	0 0	56 5	712 50	3.421 84	0 0	0 0	4.133 134	4.932 165
	00000124	56 4 S	12 1 P	2.152 29	399 11	2.619 45	0 0	123 10	0 0	123 10	345 44	2.612 61	55 3	0 0	3.012 108	5.754 163
	00000123	99 8 S	311 10 S	754 19	345 16	1.509 53	221 7	34 2 P	73 6 S	328 15	123 23	321 41	567 32	43 4	1.054 100	2.891 168
	00000122	1.837 33 S	19 1 P	88 4	0 0	1.944 38	0 0	621 13	0 0	621 13	1.015 89	2.221 52	96 18	641 8	3.973 167	6.538 218
	00000121	344 15	298 13	0 0	934 9	1.576 37	0 0	74 8	0 0	74 8	0 0	231 33	0 0	875 72	1.106 105	2.756 150
	00000120	2.657 65 S	651 28 S	3.405 70	1.678 36	8.391 199	221 7	908 38 S	73 6 S	1.202 51	2.195 206	8.806 271	718 53	1.559 84	13.278 614	22.871 864
	00000113	53 2 P	221 8 S	29 3	1.001 19	1.304 32 S	0 0	0 0	0 0	0 0	11 2 P	0 0	21 2 P	0 0	32 4 S	1.336 36
	00000112	423 18	0 0	0 0	0 0	423 18	0 0	261 5 S	34 2 P	295 7	745 71	0 0	67 8	0 0	812 79	1.530 104
	00000111	28 5 S	0 0	0 0	0 0	28 5 S	0 0	0 0	0 0	0 0	148 25 S	0 0	81 7 S	0 0	229 32 S	257 37
	00000110	504 25 S	221 8 S	29 3	1.001 19	1.755 55	0 0	261 5 S	34 2 P	295 7	904 98	0 0	169 17	0 0	1.073 115	3.123 177
	00000100	6.192 138	2.544 76	6.317 115	3.893 93	18.946 422	8.629 160	10.607 175	2.724 76	21.960 411	4.618 402	9.450 305	1.121 106	1.910 103	17.099 916	58.005 1.749

Legende:

Wert Berichtspfl. 10.000 100 **P**

Sperrvermerk (P=primär 10.000 100 **S**

Sperrvermerk (S=Sekundär) 10.000 100 **S2**

zusätzlicher Sperrvermerk (S2=sekundär ohne Schätzwertfehler)

- 194 -

Die in der Abbildung gezeigte Beispieltabelle enthält sekundäre Sperrvermerke, die bei einem vollständigen Durchlauf als positive Tabelle mit einer relativen Mindestspannweite von 2,99 bearbeitet wurde und bei der in einem zweiten Durchlauf eine zusätzliche Eingrenzung der Tabellenwerte durch einen externen Schätzfehler von 400 % berücksichtigt wurde. Beide Läufe erfolgten der einfacheren Nachvollziehbarkeit halber ohne Randschranken und ohne Übertrag von Schutzintervallen.

Das Muster der Sekundärsperrungen unterscheidet sich in beiden Fällen um nur eine zusätzliche Sekundärsperrung (Übersperrung) des Feldes (134; AAD). Zur Erklärung dieser zusätzlichen Sperrung, die gerade bei der nur positiven Tabelle auftritt, nicht aber bei der durch den 400 % Schätzfehler zusätzlich eingeengten Tabelle, genügt die Betrachtung des obersten Zeilenstreifens, bestehend aus den ersten fünf Zeilen. Die Summenzeile dieses Streifens enthält keine Sperrungen, so dass er in Bezug auf die sekundäre Geheimhaltung völlig unabhängig vom Rest der Tabelle behandelt werden kann.

Die Sekundärsperrung des Tabellenfeldes (134; AAD) entsteht durch den zum Schutze des primär geheimen Feldes (134; AAA) aufgebauten Quader {(134; AAD); (134; AAA); (131; AAD); (131; AAA)}, wenn bei einer relativen Mindestspannweite 2,99 nur die Positivität der Tabelle berücksichtigt wird: Dieser Quader mit den beiden Teilgesamtheiten von Tabellenwerten (95; 34) und (321; 256) hat dann die Spannweite 34 + 256 = 290. Bezogen auf den primär geheimen Wert ist 290/95 = 305,3 % größer als die geforderte relative Mindestspannweite von 299 %, sodass dieser Quader für eine nur positive Tabelle einen gültigen Schutzquader für das primär geheime Feld (134; AAA) darstellt. Weil kein geeigneterer Quader in dieser Untertabelle gefunden werden kann - alle anderen haben eine größere Quaderwertesumme - werden die drei anderen noch offenen Tabellenwerte dieses Quaders gesperrt. Damit ist insbesondere die Sperrung des Feldes (134; AAD) geklärt.

Die Sperrungen der Summenspalte AA werden bei der Sicherung einer nur positiven Tabelle durch den Untertabellenabgleich eingetragen; sie erzwingen auch die Sekundärsperrung des Feldes (133; AAD) im Inneren der Untertabelle aus den Spalten AAD, AAC, AAB, AAA, AA.

Um die Sekundärsperrungen in der Summenspalte AA im Falle der nur positiven Tabelle zu verstehen, betrachte man zunächst die Untertabelle der Spalten ACD, ACC, ACB, ACA und AC im obersten Tabellenstreifen. Das Einzelangabefeld (133; ACD) hat im Tabelleninneren der Untertabelle keinen Schutzquader, dessen auf seinen primär geheimen Wert bezogene Spannweite größer als die vorgegebene Mindestspannweite 2,99 wäre, sodass die Einzelangabe durch einen Quader mit den Randsummenwerten der Felder (134; AC) und (133; AC) gesichert werden muss. Der Abgleich dieser Untertabelle mit den anderen des betrachteten obersten Zeilenstreifens führt zu den Zwischensummensperrungen (134; AA) und (133; AA), den obersten sekundär geheimen Feldern in der Randsumme der ganz rechten Untertabelle unterster Aggregationsstufe.

Bei der Erklärung der geheimen Felder (133; AAD) und (131; AA) muss man davon ausgehen, dass der erste Untertabellenabgleich bereits erfolgt ist, dass also die beiden oberen Randsummenfelder (134; AA), (133; AA) bereits gesperrt sind! Dann wählt das Verfahren zur Sicherung der zuerst zu bearbeitenden - weil in der obersten Zeile stehenden - Sekundärsperrung (134; AA) das Karree {(134; AA); (134; AAA); (131; AA); (131; AAA)} und

nicht {(134; AA); (134; AAD); (133;AA); (133; AAD)}, weil die Quadersumme aufgrund der fehlenden Randschranke kleiner ist als beim zweiten Karree, denn es muss der Randwert 290 zusätzlich gesperrt werden und nicht der größere Wert 732 im Tabelleninneren. Schutzintervalle finden hier als Schätzintervalle bei der Sicherung sekundär geheimer Werte keine Berücksichtigung!

An dieser Stelle zeigt sich wieder das schon in Abschnitt 3.1.2 (2. Anmerkung) angesprochene Problem der Übertragung von Quaderspannweiten im Rahmen des Untertabellenabgleichs: Wollte man nämlich bei der Sicherung eines geheimen Wertes, z.B. des Pivots (133; ACD), mit Hilfe von Randsummensperrungen auch deren Quaderspannweite gemäß 3.2 als Schätzfehler berücksichtigen, so müsste dieser ja bei jeder Quadersicherung mit Randsummenwerten bereits vorab bekannt sein, was nicht möglich ist, weil der Abgleich erst im Nachhinein, in diesem Fall mit den beiden Werten in der Spalte AA geschieht. Dies ist ein weiteres Argument für den Aufbau vollständiger Tabellen, weil darin kein Randsummenabgleich erfolgt, sondern alle Sicherungsquader in nur einer einzigen Tabelle erstellt werden und somit jedem Quaderwert eine einheitliche Spannweite zukommt.

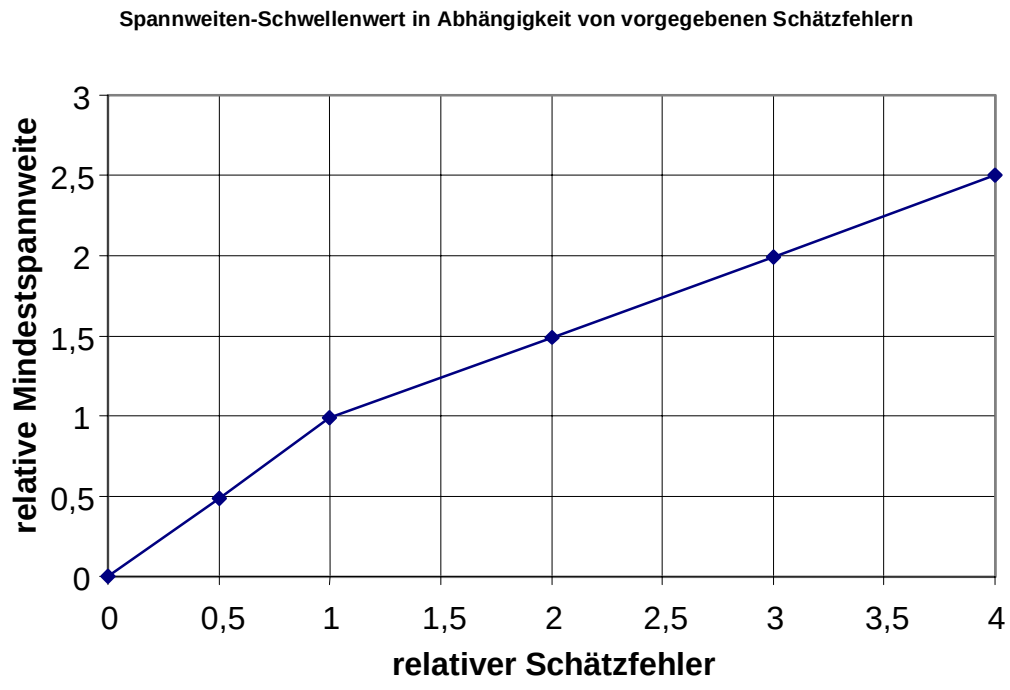
Als nächstes zu sicherndes sekundär gesperrtes Randsummenfeld steht jetzt (133; AA) an. Jetzt erst muss das bis dahin noch offene Feld (133; AAD) zum Aufbau z. B. des Sicherungsquaders {(133; AA); (133; AAD); (131; AA); (131; AAD)} gesperrt werden. Der Abgleich durch Bildung entsprechender Sicherungsquader in der Untertabelle höherer Zeilenaggregation, bestehend aus den Spalten (AC, AB, AA mit Randsumme A) ergibt dann die Sekundärsperrung (131; AC) und die „Gegensperrung“ (131; ACC) im Inneren der ganz linken Untertabelle unterste Aggregationsstufe - selbstverständlich wieder als Elemente entsprechender Sicherungsquader.

Die bisherigen Betrachtungen der obigen Beispieltabelle bezogen sich allesamt auf eine nur positive Tabelle, d.h. auf eine Tabelle, von der der externe Nutzer nur weiß, dass alle Werte nicht negativ sind und über die er keine anderen Angaben hat, wie z.B. Schätzintervalle, die die Tabellenwerte überdecken. Lässt man jetzt auch solche externen Schätzintervalle zu, wie hier in Form eines Schätzfehlers von 400 %, so kann der primär geheime Wert im Feld (134; AAA) nicht mehr durch das Karree {(134; AAA); (134; AAD); (131; AAA); (131; AAD)} im Inneren der rechten Untertabelle unterster Aggregation gesichert werden, weil jetzt gemäß Punkt 3.2.2 Formel (10) neben den kleinsten Werten der beiden Quaderteilgesamtheiten, 34, 256, auch noch der kleinste externe Schätzfehler des Quaders, hier 400 % * 34/100 = 136 zu berücksichtigen ist. Die Quaderspannweite ist jetzt min(136; 34) + min(136; 256) = 34 + 136 = 170 bzw. die relative Spannweite des primär geheimen Wertes 170/95 = 1,79 also kleiner als die relative Mindestspannweite 2,99; der Quader ist als Sicherungsquader abzulehnen. Stattdessen bleibt nur der Quader {(134; AA); (134; AAA); (131; AA); (131; AAA)} als Sicherungsquader übrig, der gemäß (10) und (11) folgende Spannweite hat, min(380; 95) + min(380; 256) = 95 + 256 = 351 oder die relative Spannweite des primär geheimen Wertes 351/95 = 3,695, die deutlich größer als die vorgegebene relative Mindestspannweite von 2,99 ist. Dabei wurde der kleinste externe Schätzfehler des Quaders nach 4 * 95 = 380 berechnet.

Die anderen sekundär geheimen Werte der ganz rechten Untertabelle des obersten Zeilenstreifens, (133; AA), (133; AAD) sowie (131; AAD), werden wieder durch den Untertabellenabgleich mit der ganz linken Untertabelle eingetragen und zwar auf genau die gleiche Weise wie beim Abgleich ohne externe Schätzintervalle , denn bei der Sicherung sekundär geheimer Werte bleibt, wie oben bemerkt, der Intervallschutz unberücksichtigt. Das Sperrmuster des

obersten Zeilenstreifens ist damit aufgeklärt und auch die Eigenart, dass eine Sicherung mit Berücksichtigung von externen Schätzintervallen u.U. zu weniger Sperrungen führt als in einer positiven Tabelle ohne Angaben über Schätzintervalle, denn das fragliche Feld (134; AAD) wird bei Sicherung der positiven Tabelle ohne externe Schätzfehler benötigt, bei Sicherung mit Berücksichtigung externer Schätzfehler von 400 % aber nicht!

Eine wesentliche Besonderheit der Sicherung von Tabellen mit externen Schätzintervallen ist - wie bereits bemerkt - , die Tatsache, dass man bei nicht zu großen Schätzfehlern relative Mindestspannweiten vorgeben kann, die durch kein Geheimhaltungsverfahren realisiert werden können. Dies zeigt sich in der grafischen Darstellung der Anzahl der Sekundärsperrungen in Abhängigkeit von der relativen Mindestspannweite mit externem Schätzfehler als Parameter durch die Abbrüche der Kurven bei zu großen relativen Mindestspannweiten. In der folgenden grafischen Darstellung sind für den steuerbaren Umsatz NRW 1994 die Spannweitenschwellenwerte, die relative Mindestspannweite, die gerade noch nicht zum Abbruch des Geheimhaltungslaufs führt, in Abhängigkeit vom relativen externen Schätzfehler aufgetragen, um bei gegebenem relativen Schätzfehler die gerade noch zulässige relative Mindestspannweite auswählen zu können. Umgekehrt erhält man durch diese Darstellung auch einen Eindruck, wie groß der externe Schätzfehler mindestens sein muss, damit beispielsweise eine relative Mindestspannweite größer als 1, wie sie der Schutz dominierender Werte erfordert (vergleiche dazu 4.2 in „Statistische Analysen und Studien NRW“, 3/2000), noch gesichert werden kann.



Erwartungsgemäß nimmt der gerade noch zu realisierende Intervallschutz in Gestalt der vorgebbaren relativen Mindestspannweite mit abnehmender Vorinformation, d.h. mit zunehmendem relativen Schätzfehler bis 100% in derselben Weise zu, wie der relative Schätzfehler und zwar mit einer Steigerung von 1 : 1. Schätzfehlergrenzen über 100% wirken sich bei positiven Tabellen wie im vorliegenden Fall nur noch auf die obere Schätzintervallgrenze

aus, sodass der weitere Anstieg der relativen Mindestspannweite nur noch halb so groß ausfällt wie unterhalb von 100%.

Man sieht, dass bei einer aus Sensitivitätsgründen zu fordernden relativen Mindestspannweite größer als 1, die mit den Daten in der Regel gut vertrauten Tabellennutzer die Tabellenwerte nicht einmal mehr bis auf +/- 100 % genau eingrenzen können dürfen, damit eine Sicherung mit Sensitivitätsschutz überhaupt noch möglich ist. Wenn man aber Schätzintervalllängen als ursprüngliche Maße für die Empfindlichkeit geheimer Tabellenwerte ansieht, kommt man mit relativen Mindestspannweiten von beispielsweise 30 % aus. Dann kann beim externen Tabellennutzer aber schon ein recht genaues Datenwissen vorausgesetzt werden, er mag die Daten noch vor der Veröffentlichung der Tabelle lt. Darstellung bis nahezu 30% genau festlegen können und trotzdem ist eine sekundäre Geheimhaltung mit Intervallschutz noch durchführbar.

A.4 Quaderverfahren in vollständigen Tabellen

A.4.1 Aufgestockte verkürzte Umsatzsteuerstatistik NRW 1994

Zur Demonstration einer Anwendung des Quaderverfahrens auf eine aufgestockte vollständige Tabelle von Realdaten wurde die Umsatzsteuerstatistik für Nordrhein-Westfalen von 1994 regional vom Land über die Regierungsbezirke bis zu den Kreisen und kreisfreien Städte gegliedert, von der Wirtschaftssystematik wurden zunächst drei Aggregationsstufen aufgenommen, die Unterabschnitte, die Abschnitte und die Summe (siehe die beigefügte Statistik über die Aufstockung: Zur vollständigen 4-dimensionalen Tabelle ...). Die so erhaltene verkürzte Umsatzsteuerstatistik ist hinsichtlich ihrer Gliederungsstruktur mit der unter 1.2.2 eingeführten Beispieltabelle vergleichbar; trotz ihrer starken Verdichtung umfasst sie aber immer noch mehr als zehnmal so viele Tabellenfelder wie die Beispieltabelle. Die dimensionsaufgestockte Umsatzsteuerstatistik ist mit 23.040 Tabellenfeldern eine selbst in dieser verkürzten Form noch recht umfangreiche vierdimensionale Tabelle - die aufgestockte Beispieltabelle besteht dagegen aus nur 480 Tabellenfeldern. Durch die Dimensionsaufstockung wird die Umsatzsteuertabelle um mehr als das 8-fache erweitert, die Beispieltabelle aber nur um das Doppelte.

Als vierdimensionale Tabelle ist die aufgestockte Umsatzsteuerstatistik mit den derzeit in der Anwendung befindlichen EDV-Programmen ab GHQUAR.4.1 ohne weiteres zu bearbeiten, da bei diesen Programmen die dimensionsbedingte Anwendungsgrenze erst bei sieben Dimensionen liegt; es muss allerdings eine entsprechende Erweiterung des von dem EDV-Programm zu belegenden Arbeitsspeichers vorgenommen werden. Die Beschränkung auf „nur“ vier Dimensionen wird bei der hier vorliegenden feinen Gliederung bereits durch die zu erwartenden großen Rechenzeiten auferlegt. Es sei daran erinnert, dass der Tabellenumfang den Rechenzeitaufwand zwar nur quadratisch bestimmt, die Tabellendimension aber exponentiell in die Anzahl der Rechenoperationen eingeht (vergleiche Absatz 2.2.2). Außerdem wurde vereinfachend auf Intervallschutz verzichtet. Wie vorhergehende Untersuchungen gezeigt haben (vergleiche A.1) wird die Rechenzeit durch Berücksichtigung von Intervallschutz nicht

wesentlich erhöht, und der hier aufzuzeigende Einfluss der Tabellenumstrukturierung auf die Verteilung der Sekundärsperrungen bleibt auch im Falle der Tabellensicherung mit Intervallschutz in derselben Weise wirksam.

Trotz des großen Unterschiedes zwischen dem Tabellenumfang der Umsatzsteuer und dem der Beispieltabelle weisen beide Tabellen - und zwar sowohl in zweidimensionaler als auch in der aufgestockten vierdimensionalen Form - etwa die gleichen Anzahlen von Sekundärsperrungen aus und das obwohl in die Umsatzsteuerstatistik etwa 40mal so viele Primärsperrungen eingetragen sind wie in die Beispieltabelle. Eine Erklärung dieses Phänomens bietet die unterschiedliche Verteilung der Primärsperrungen über die beiden Gesamttabellen: Während die Primärsperrungen über die Beispieltabelle nahezu gleichmäßig verteilt sind, bilden sich in der realen Tabelle auf Grund „strukturschwacher“ Bereiche Anhäufungen von primär geheimen Werten, weil die in solchen Bereichen vorliegenden geringeren Anzahlen von Berichtenden häufiger zu Tabellenfeldern mit weniger als drei Merkmalsträgern führen, die primär geheim gehalten werden müssen. In diesen Bereichen sind Primärsperrungen oftmals von vielen anderen primär geheimen Werten umgeben, so dass sie sich gegenseitig schützen.

Zur vollständigen 4-dimensionalen Tabelle der verkürzten Umsatzsteuerstatistik NRW 1994[14]

Bestehend aus den Gliederungen :	regional -	Kreise und kreisfreie Städte
		Regierungsbezirke
		Land NRW
	wirtschaftl. -	Unterabschnitt
		Abschnitt
		Summe

Anzahl der gesamten/besetzten Tabellenfelder:	23.040 / 11.087
Anzahl der Dummies in den besetzten Tabellenfeldern	8.535
Anzahl der Primärsperrungen/Einzelberichtspflichtigen	357 / 81
Anzahl der gesamten/besetzten Tabellenfelder (2-dim.)	2700 / 2552

Sicherung ohne Randschranke:

[14] Die Auswertungen wurden auf dem IBM-Großrechner des LDS NRW (ES 9000 – OS 390) durchgeführt.

2-dimensional :	Anzahl der Sekundärsperrungen	26
	CPU-Zeit in Sekunden	~1 (normales Untertabellenverfahren)
4-dimensional:	Anzahl der Sekundärsperrungen	25
	CPU-Zeit in Sekunden	30

Sicherung mit Randschranken:

2-dimensional :	Anzahl der Sekundärsperrungen	26
	CPU-Zeit in Sekunden	~1 (normales Untertabellenverfahren)
4-dimensional:	Anzahl der Sekundärsperrungen	26
	CPU-Zeit in Sekunden	30

Um das enorme Anwachsen von Tabellenumfang und Rechenzeit zu verdeutlichen, das im Allgemeinen mit einer Verfeinerung der Tabellengliederung einhergeht, wurde der oben vorgestellte verkürzte steuerbare Umsatz noch um nur eine Aggregationsstufe in wirtschaftlicher Gliederung erweitert: Die wirtschaftliche Gliederung überdeckt nun die 4 Gliederungsstufen (aufsteigend) Abteilung, Unterabschnitt, Abschnitt und Summe (siehe Übersicht „Zur vollständigen 5-dimensionalen Tabelle ...). Die dimensionsaufgestockte Tabelle ist damit eine fünfdimensionale und umfasst mit Dummy-Werten 138.240 Datensätze bzw. Tabellenfelder, die ursprüngliche (nicht aufgestockte) zweidimensionale Tabelle enthält dagegen nur 6.180 Tabellenfelder. Erbrachte die Aufstockung der zweidimensionalen zur vierdimensionalen Umsatzsteuertabelle noch eine 8-fache Erweiterung des Tabellenumfangs, so bewirkt die Aufstockung der nur um eine einzige Aggregationsstufe erweiterten zur vollständigen fünfdimensionalen Tabelle bereits eine Tabellenvergrößerung um mehr als das Zweiundzwanzigfache.

Zur vollständigen 5-dimensionalen Tabelle der verkürzten Umsatzsteuerstatistik NRW 1994

Bestehend aus den Gliederungen :	regional -	Kreise und kreisfreie Städte
		Regierungsbezirke
		Land NRW
	wirtschaftl. -	Abteilung
		Unterabschnitt

	Abschnitt	Summe
Anzahl der gesamten/besetzten Tabellenfelder:		138.240 / 119.346
Anzahl der Dummies in den besetzten Tabellenfeldern		113.817
Anzahl der Primärsperrungen/Einzelberichtspflichtigen	617 /	227
Anzahl der gesamten/besetzten Tabellenfelder (2-dim.)	6180 /	5529

Sicherung ohne Randschranke:

2-dimensional :	Anzahl der Sekundärsperrungen	470	
	CPU-Zeit in Sekunden	1	(normales Untertabellenverfahren)
5-dimensional:	Anzahl der Sekundärsperrungen	414	(mit Sperrungen in den höchsten Aggregationsstufen – Eckfeldsperrung)
	CPU-Zeit in Sekunden	750	

Sicherung mit Randschranken:

2-dimensional :	Anzahl der Sekundärsperrungen	462	
	CPU-Zeit in Sekunden	1	(normales Untertabellenverfahren)
5-dimensional:	Anzahl der Sekundärsperrungen	464	(mit Sperrungen in höchster Aggregation – Eckfeldsperrung)
	CPU-Zeit in Sekunden	750	

Noch wesentlich drastischer fällt allerdings die Steigerung der Rechenzeit bei der sekundären Geheimhaltung der vollständigen fünfdimensionalen gegenüber der vierdimensionalen Tabelle des steuerbaren Umsatzes aus:

Für die Bearbeitung der vollständigen vierdimensionalen Umsatzsteuerstatistik wurde 30mal so viel Rechenzeit (CPU-Zeit) benötigt wie für die unaufgestockte zweidimensionale Tabelle bei Sicherung mit iterativem Untertabellenabgleich, für die vollständige fünfdimensionale Tabelle aber 750 mal soviel CPU-Zeit wie für die zugehörige zweidimensionale Ausgangstabelle! Die erforderliche Rechenzeit ließe sich jedoch beträchtlich verringern, wenn man beispielsweise an Stelle der Gesamttabelle nur die durch Zwischensummen ohne Sperrungen abgegrenzten Tabellenteile einzeln behandeln könnte. Solche Tabellenteile sind nach ihren individuellen Dimensionsaufstockungen nicht nur hinsichtlich ihres Tabellenumfangs wesentlich reduziert, sie weisen auch eine kleinere Tabellendi-

mension auf und sind infolge dessen sehr viel schneller zu bearbeiten als die vollständige Gesamttabelle, was der Vergleich der vierdimensionalen mit der fünfdimensionalen Umsatzsteuerstatistik besonders deutlich macht. Auf diese Möglichkeit der Rechenzeitverkürzung wurde bereits im Abschnitt 6.2.2.2 ausdrücklich hingewiesen; sie wurde mit weiteren rechenzeitsparenden Methoden in Abschnitt 7. ausführlich dargestellt.

Das wohl Bemerkenswerteste an den vorliegenden Ergebnissen bei der Sicherung der aufgestockten verkürzten Umsatzsteuerstatistik ist die schon bei der Bearbeitung der Beispieltabelle nach deren Dimensionsaufstockung beobachtete Abnahme von Sekundärsperrungen gegenüber der mit Untertabellenabgleich gesicherten zweidimensionalen Tabelle - zumindest bei Bearbeitung ohne Setzen von Randschranken. Dass dieser Gewinn von sperrvermerkfreien Tabellenfeldern nicht allein auf die Gesamtsicht der Tabelle bei der Quaderauswahl zurückzuführen ist, zeigt die jeweils zweite Auswertung mit gesetzten Randschranken für jede Gliederung im Aufstockungsfall: Die Verhinderung von Randsperrungen erhöht die Anzahl der Sekundärsperrungen, d.h. eine Verringerung von Sekundärsperrungen in einer aufgestockten Tabelle gegenüber einer mit Untertabellenabgleich gesicherten kann u.U. auch durch vermehrte Sperrungen in die Randsummen verursacht worden sein.

A.4.2 Quaderverfahren in fiktiven vollständigen Tabellen

In diesem Abschnitt wird das Laufzeit- und Sperrverhalten der zuletzt im LDS NRW entwickelten Version des Quaderverfahrens zur Sicherung geheimer Tabellenwerte mit vollständigen Quadern (realisiert durch QUIT) anhand einer simulierten Umsatzstatistik-Tabelle untersucht. Ausgangsmaterial sind die 94-er Daten des regional und nach wirtschaftlicher Systematik gegliederten steuerbaren Umsatzes (vgl. A.1), deren Verteilung zur Monte-Carlo-Simulation eines synthetischen Datenbestandes genutzt wurde, um nicht sensible Originaldaten über längere Zeit nachhalten zu müssen (was sich aus Datenschutzgründen verbietet). Dass diese Simulation die Verhältnisse des erhobenen Bestandes recht gut wiedergibt, zeigt u.A. der direkte Vergleich der entsprechenden in diesem Abschnitt vorgelegten Daten mit denen des vorhergehenden Abschnitts.

In nachstehender Übersicht sind die Ergebnisse der Geheimhaltungsläufe mit dem EDV-Programm QUIT für vier verschiedene Gliederungstiefen den entsprechenden Ergebnissen von GHQUAR- und GHMITER-Anwendungen gegenübergestellt (Stand 26.02.2003). Mit GHQUAR wurden real aufgestockte, d.h. auf externem Datenträger vorhandene vollständige Tabellen bearbeitet. GHMITER „sichert“ die gleichen Tabellen ohne Vervollständigung durch Aufstockung, d.h. mit Untertabellenabgleich. Das EDV-Programm QUIT nutzt die partielle dynamische Dimensionsaufstockung in Gestalt der unter 7. beschriebenen Bearbeitung fiktiver Tabellen. Die Sicherungsläufe wurden ohne Intervallschutz und ohne Setzen von Randschranken durchgeführt.

Ausgangsmaterial für die Beurteilung : simulierte Umsatzsteuer von 1994 / Auszüge/ohne Randschranke[15]

1) bestehend aus : regional - Kreise und kreisfreie Städte, Regierungsbezirke und Land NRW

[15] Alle Auswertungen wurden auf PC (Pentium / 400 128 MB) unter dem Betriebssystem NT4 durchgeführt.

wirtschaftl. - Unterabschnitte, Abschnitte und Summe

Anz. der besetzten/gesamten Tabellenfelder	2.552 / 2.700
Anzahl Primärsperrungen/Einzelwerte	217 / 81
Gesamtfeldanzahl im Aufstockungsfall	23.040

2) bestehend aus : regional - Kreise und kreisfreie Städte, Regierungsbezirke und Land NRW

wirtschaftl. - Abteilungen, Unterabschnitte, Abschnitte und Summe

Anz. der besetzten/gesamten Tabellenfelder	5.529 / 6.180
Anzahl Primärsperrungen/Einzelwerte	617 / 227
Gesamtfeldanzahl im Aufstockungsfall	138.240

3) bestehend aus : regional - Kreise und kreisfreie Städte, Regierungsbezirke und Land NRW

wirtschaftl. - Unterabteilungen, Abteilungen, Unterabschnitt, Abschnitt und Summe

Anz. der besetzten/gesamten Tabellenfelder	15.482 / 19.380
Anzahl Primärsperrungen/Einzelwerte	3149 / 1417
Gesamtfeldanzahl im Aufstockungsfall	1.382.400

4) bestehend aus : regional - Gemeinden, Kreise u. kreisfr. Städte, Regierungsbez. und Land NRW

wirtschaftl. - Unterabteilungen, Abteilungen, Unterabschnitte, Abschnitte und Summe

Anz. der besetzten/gesamten Tabellenfelder	75.450 / 139.859
Anzahl Primärsperrungen/Einzelwerte	20.849/ 15.741
Gesamtfeldanzahl im Aufstockungsfall	28.569.600

Testbearbeitung mit GHQUAR und real aufgestocktem Material

Zu 1) Anzahl der sekundären Sperrungen : 103

Wertsumme der Sperrungen : 2.922.460.000.000 oder 15,38 % des Gesamtmengenwertes

CPU-Zeit : ca. 15 Sek.

Zu 2) Anzahl der sekundären Sperrungen : 465

Wertsumme der Sperrungen : 6.280.760.000.000 oder 24,79 % des Gesamtmengenwertes

CPU-Zeit : ca. 406 Sek.

Zu 3 und 4) Wegen der enormen Größe der aufgestockten Tabelle ist die Bearbeitung mit GHQUAR nicht ohne Programmierer-Eingriffe in das EDV-Programm durchführbar. Außerdem verhindern die großen Rechenzeiten die Durchführung solcher Geheimhaltungsläufe.

Testbearbeitung mit GHMITER

Zu 1) Anzahl der sekundären Sperrungen : 102

Wertsumme der Sperrungen : 25.500.298.598 oder 0,13 % des Gesamtmengenwertes

CPU-Zeit : ca. 3 Sek.

Zu 2) Anzahl der sekundären Sperrungen : 460

Wertsumme der Sperrungen : 169.853.445.438 oder 0,67 % des Gesamtmengenwertes

CPU-Zeit : ca. 7 Sek.

Zu 3) Anzahl der sekundären Sperrungen : 1609

Wertsumme der Sperrungen : 575.699.902.517 oder 1,82 % des Gesamtmengenwertes

CPU-Zeit : ca. 19 Sek.

Zu 4) Anzahl der sekundären Sperrungen : 8.258

Wertsumme der Sperrungen : 881.696.568.266 oder 2,51 % des Gesamtmengenwertes

CPU-Zeit : ca. 154 Sek.

Testbearbeitung mit QUIT (Stand 25.02.2003)

Zu 1) Anzahl der sekundären Sperrungen : 95

	Wertsumme der Sperrungen	:	38.900.101.348 oder 0,20 % des Gesamtmengenwertes
	CPU-Zeit	:	ca. 7 Sek.
Zu 2)	Anzahl der sekundären Sperrungen	:	465
	Wertsumme der Sperrungen	:	255.231.281.459 oder 1,01 % des Gesamtmengenwertes
	CPU-Zeit	:	ca. 14 Sek.
Zu 3)	Anzahl der sekundären Sperrungen	:	1630
	Wertsumme der Sperrungen	:	1.102.605.762.500 oder 3,48 % des Gesamtmengenwertes
	CPU-Zeit	:	ca. 82 Sek.
Zu 4)	Anzahl der sekundären Sperrungen	:	11.389
	Wertsumme der Sperrungen	:	12.572.930.550.645 oder 35,74% des Gesamtmengenwertes
	CPU-Zeit	:	ca. 1Std 56min 3sec
			Eckfeldsperrungen in höchster Ebene (LAND/SUMME)

Von ausschlaggebender Bedeutung für den praktischen Einsatz des Quaderverfahrens ist sein Laufzeitverhalten. Das gilt in hohem Maße auch für das EDV-Programm QUIT. Daher wird diesem Aspekt im Folgenden besondere Beachtung geschenkt.

In den oben aufgeführten Testergebnissen zeigt das Laufzeitverhalten von QUIT den erwarteten deutlichen Anstieg der Rechenzeit mit zunehmender Verfeinerung der Tabellengliederung: Bei 2700 Tabellenfeldern (zu1) werden 7 Sekunden (s) benötigt, das sind 2,6 Millisekunden (ms) pro Tabellenfeld. Bei 6180 Feldern (zu 2) sind es 14 s oder 2,3 ms / Feld. Entsprechend hat man (zu 3) 82 s / 19380 Felder = 4,2 ms / Feld und (zu 4) 6963 s / 139859 Felder = 49.8 ms / Feld. Bezieht man die Laufzeit auf das Anzahlquadrat der Tabellenfelder, ergeben sich die Werte 0,96; 0,37; 0,22; 0,36 in Mikrosekunden pro Anzahlquadrat der Tabellenfelder.

D.h. bei QUIT ist die Rechenzeit pro Anzahlquadrat der Tabellenfelder in dem hier betrachteten Bereich nahezu konstant. Bei GHMITER findet man hingegen schon für die Laufzeiten (ms) pro Tabellenfeld konstante Werte, 1,11; 1,13; 0,98; 1,10 , also einen linearen Zusammenhang von Laufzeit und Tabellenumfang. Dass bei QUIT der erste Wert nach oben abweicht, lässt sich mit den gegenüber GHMITER größeren Rüstzeiten erklären, die zu einem von den Tabellenparametern unabhängigen und mit kleinen Rechenzeiten im Sekundenbereich vergleichbaren Sockelbetrag an Zeit führen.

Der bei GHMITER beobachtete lineare Zusammenhang der Rechenzeit mit dem Tabellenumfang lässt sich durch die Zerlegung der Gesamtrechenzeit in die Untertabellenrechenzeiten plausibel machen, wenn man davon ausgeht,

dass die bei jedem Erweiterungsschritt neu hinzukommenden Untertabellen im Mittel etwa genau so groß sind wie

die bereits vorhandenen und dies auch als Näherung für die Untertabellenrechenzeiten gilt. Bei diesem Ansatz ist der Gesamtumfang und die Gesamtrechenzeit der Tabelle proportional zur Anzahl der Untertabellen, also die Gesamtrechenzeit der Tabelle proportional zu ihrem Gesamtumfang.

Die Laufzeit von QUIT ergibt sich aus der Summe der Zeiten, die für die Sicherung jedes einzelnen primär geheimen Wertes aufzuwenden sind. Sie lässt sich demnach als Produkt aus der Anzahl der Primärfälle und der mittleren Bearbeitungszeit eines Primärfalls darstellen. Da die beobachtete Laufzeit von QUIT mit dem Tabellenumfang annähernd quadratisch zunimmt und die Anzahl der Primärfälle selbst schon mit der Anzahl der Tabellenfelder ansteigt, besteht nur noch eine „schwache“ Abhängigkeit der mittlere Bearbeitungszeit je Primärfall vom Tabellenumfang. Dies deutet darauf hin, dass gerade die besonders zeitaufwendigen hochaggregierten Primärsperrungen durch die unter Abschnitt 7 beschriebene Strangbehandlung weitgehend haben „entschärft“ werden können.

Die recht grobe Behandlung des Zusammenhangs von Rechenzeit und Tabellenumfang wird bei GHMITER durch die unterschiedlichen Untertabellenumfänge bzw. bei QUIT durch die unterschiedlichen Dimensionen und Umfänge der temporär aufgestockten Teiltabellen nahegelegt. Eine Anwendung der unter 2.2.2.1 angegebenen Rechenformel bleibt im Falle hierarchischer Gliederung allein einer real aufgestockten Tabelle vorbehalten, wie sie unter A.4.1 und in diesem Abschnitt unter „Testbearbeitung mit GHQUAR ...“ vorgestellt worden ist.

Um die Rechenzeit für die hier nicht mehr ausgewertete real aufgestockte 6-dimensionale Tabelle abzuschätzen, kann man R als Rechenzeit (anstelle der Anzahl von Rechenschritten) interpretieren, die unbekannten Segmentlängen m_i , i = 1, ..., n aus der Gesamtzahl der Tabellenfelder durch die einheitliche Länge $T^{1/n}$ abschätzen und aus den beiden bearbeiteten aufgestockten Tabellen die Konstante a bestimmen. Man erhält so die Tabelle

	n	R[s]	N_g	T	$M=(T^{1/n}-1)^n$	2^n	$a = R / (N_g * M * 2^n)$
zu 1	4	15	298	23040	16422,1	16	0,1916 µs
zu 2	5	406	844	138240	84513,2	32	0,1779 µs

Mit dem aus dieser Tabelle zu entnehmenden Zeitwert einer elementaren Rechenoperation von a = 0,185µs, der Anzahl N_g = 4566 primär geheimen Werte in der 6-dimensionalen Tabelle insgesamt und der Anzahl T = 1 382 400 Tabellenfelder der aufgestockten Tabelle (M = 760775) ergibt sich für die Rechenzeit der aufgestockten 6-dimensionalen Tabelle R = 41129 s > 11 h. Dieser Wert ist zu vergleichen mit dem der QUIT-Auswertung, die für die dritte Tabelle ganze 82 s gedauert hat.

Um das Laufzeitverhalten der beiden für die Anwendung bereitgestellten EDV-Programme GHMITER und QUIT zu vergleichen, ist noch folgende Tabelle angefügt:

Laufzeiten von GHMITER und QUIT in Abhängigkeit vom Tabellenumfang

Felderzahl / EDV-Programm	zu1: 2700	zu2: 6180	zu3: 19380	zu4: 139859
GHMITER	3 s	7 s	19 s	154 s
QUIT	7 s	14 s	82 s	6963 s

Man sieht, Tabellen mittlerer Größe bis etwa 50 000 Tabellenfelder können ohne Weiteres mit QUIT gesichert werden. Bei größeren Tabellen empfiehlt sich im Falle nicht zu sensibler Daten die Anwendung von GHMITER oder, bei höherem Sicherheitsanspruch, die Aufteilung in mittlere Tabellen, die als Pool voneinander abhängiger Tabellen mit QUIT gesichert werden können.

Wie in dieser Dokumentation einleitend angemerkt, sind Sekundär-Geheimhaltungsverfahren so zu konzipieren, dass durch Sperrungen auf den niedrigeren Aggregationsniveaus die höheren Niveaus veröffentlicht werden können. Wie weit dies bei QUIT gelungen ist, zeigt ein Vergleich mit GHMITER im Falle von Untertabellenabgleich und mit GHQUAR bei Anwendung des „starren“ Quaderverfahrens auf real aufgestockte Tabellen:

Anteil gesperrter Werte an der Gesamtsumme bei GHMITER, bei GHQUAR und bei QUIT

Felderzahl / EDV-Programm	zu1: 2700	zu2: 6180	zu3: 19380	zu4: 139859
GHMITER	0,13%	0.67%	1,82%	2,51%
GHQUAR	15,38%	24,79%	- - -	- - -
QUIT	0,20%	1,01%	3,48%	35,74%

Man sieht auch hier: Tabellen mittlerer Größe, d.h. mit bis zu ca 50 000 Tabellenfeldern und nicht zu feiner hierarchischer Gliederung sind mit QUIT zu sichern, ohne im Vergleich zu einer Sicherung mit GHMITER große Zunahmen an Summensperrungen hinnehmen zu müssen. Ein Vergleich des Sperrverhaltens von GHQUAR mit GHMITER und QUIT fällt für eine Bearbeitung mit starrer Quaderstruktur recht ernüchternd aus: Selbst für die hier präsentierten vergleichsweise groben Gliederungen übersteigt der Anteil der gesperrten Summen an der Gesamtsumme die entsprechenden Vergleichswerte von GHMITER und QUIT um eine Größenordnung.

Zusammenfassend kann gesagt werden, dass das Verfahren der dynamischen Tabellenaufstockung mit Strangbearbeitung und Quaderdeformation, beurteilt anhand der Erfahrungen mit QUIT, bei Veröffentlichungstabellen mit bis zu 50 000 Feldern sowohl hinsichtlich seines Laufzeitverhaltens wie auch in Bezug auf die Öffnung von Summenwerten für eine hinreichende Sicherung primär geheimer Werte geeignet ist. Bei größeren Tabellen kann eine Aufteilung der Gesamttabelle in kleinere Tabellenteile, die dann mit QUIT als Pool abhängiger Tabellen zu sichern sind, insbesondere dann sinnvoll sein, wenn es sich um besonders sensibles Datenmaterial mit hohem Sicherheitsanspruch handelt.

A.5 Übersicht über Anwendungsmöglichkeiten

Die Einsatzmöglichkeiten des Quadersicherungsprinzips für den Schutz geheimer sensibler Tabellendaten sind äußerst vielgestaltig. Sie werden hauptsächlich bestimmt durch die Faktoren Tabellentyp, Vorinformation über die Tabellenwerte, Tabellenorganisation, Bewertung der von der Geheimhaltung betroffenen Tabellenwerte, die durch den Faktor „Justierung“ eingetragen wird, Art und Grad der Sicherung.

Diese Faktoren können bei der Bearbeitung von Geheimhaltungsproblemen nicht vollständig unabhängig voneinander realisiert werden. Hat man es beispielsweise mit einander überlappenden Statistiktabellen zu tun, so wird man die zu bearbeitenden Einzeltabellen nach der in Kapitel 6.1 beschriebenen Weise zusammenführen und bei der Geheimhaltung entsprechend organisieren (eine spezielle Kategorie des Faktors Tabellenorganisation), dabei ist der Grad der Sicherung i.A. kein hinreichender sondern nur ein notwendiger. Um die überlappenden Tabellen nach Möglichkeit weitgehend zu entkoppeln, wird man sich der Justierung durch Setzen von Randschranken bedienen, wodurch der Faktor Tabellenorganisation mit dem der Justierung verknüpft ist.

Die gegenseitige Abhängigkeit der Faktoren schränkt die bestehenden Auswahlmöglichkeiten der Faktorkategorien also weitgehend ein. Um die Entscheidung zu erleichtern, welcher Satz von Faktorausprägungen bei einem vorgelegten Geheimhaltungsproblem der geeignete ist, wurde die nachstehende Übersichtstabelle angefügt. Sie enthält die in Betracht kommenden Faktoren mit ihren Kategorien, die Kapitelnummern, unter denen diese Faktoren abgehandelt werden, und Bemerkungen mit anwendungsrelevanten Hinweisen.

Quaderverfahren für n-dimensionale Tabellen					
Faktor	Faktorkategorie/Kapitel/Bemerkung				
Tabellentyp	Wertetabelle ohne Zwischensummen	Wertetabelle mit Zwischensummen	überlappende Tabellen	Zeitreihen-Tabellen	Kontingenztabellen
	1.1, 2. u. 3.	1.2, 2., 3. und / oder 6.2	6.1	3.2.3.1 und 5.3.2	Einführung
	ohne Zwischensummen gegeben	Untertabellenhierarchie oder Aufstockung	Randschranken einsetzen	externe Gewichtung einsetzen	Ersetzung der Werte durch Fälle
Vorinformation	Tabelle enthält positive und negative Werte	Tabelle enthält nur positive Werte	für Tabellenwerte existieren Schätzintervalle		
	5.1.2.1	3.1	3.2		
	Nullwerte haben keine Sonderstellung	Nullen nur in einer Quaderteilgesamtheit	zu kleine Schätzintervalle verhindern Geheimhaltung		
Tabellenorganisation	Untertabellen-Hierarchie mit Abgleich	Vervollständigung durch Aufstockung	Zusammenführung überlappend. Tab. in gemeinsam. Datenbest.		
	1.2	6.2	6.1		
	Extraktion aller Tabellenteile gleicher Dimension mit Rand- ohne Zwischensummen	(temporärer) Aufbau von Tabellen(-Teilen) ohne Zwischensummen	Tabellenverschränkung oder übergeordneter Tabellenraum		
Justierung	Setzen von Randschranken (interne Gewichtung)		externe Gewichtung z.B. werte-, fallzahl-, positionsbezogen		
	5.2.1		5.3		
	dimensionsabhängiger Schrankenwert vermeidet oder fördert Randsperrungen		Randwertgewichtung nicht durch Schrankensetzung, sondern nur extern möglich		
Art der Sicherung	Sekundärsperrungen gegen		Verfälschungen durch		
	eindeutige Rückrechnung	zu genaue Rückrechnung	Umbuchungen	Zufallsfelder ε	
	2.	3.	4.1	4.1	
	ohne Intervallschutz	mit Intervallschutz	Austausch von Fällen innerhalb des Quaders	gerade indiziert: + ε ungerade indiziert: - ε	
Grad der Sicherung	bei Tabellenabgleich nur notwendiger, kein hinreichender Schutz		hinreichender Schutz nur bei Tabellen ohne Zwischensummen		
	6.2.1		6.2.2 und 7.		
	Beispiele: Untertabellenabgleich, Abgleich von überlappenden Tabellen		zu erreichen durch Vermeidung von Summensperrungen und durch Dimensionsaufstockung		

B Begriffsbestimmungen

Begriff	Erläuterung	Abschnitt
Aggregat(wert)	Ein Aggregat ist ein Datenwert, der durch reine Summierung ermittelt wird - entweder aus dem Einzelmaterial durch -- Fallzählung (+1), sog. Fallzahl, oder -- Addition von Einzelwerten (+EF), sog. Summenwert, - oder durch Summierung bereits ermittelter Aggregate, wie sie z.B. in Tabellen als Randsummen vorkommen.	z.B. Einführung
Aggregationsniveau	Höhe der Gliederungsstufe in der Gliederungshierarchie, die ein Maß ist für die Vergröberung (Zusammenfassung, Verdichtung) durch die Summierung in der betreffenden Gliederung. Z.B. belegen in der regionalen Gliederung von NRW die Gemeinden das unterste, die Kreise und kreisfreien Städte das nächst höhere, die Regierungsbezirke das zweihöchste und das Land das höchste Aggregationsniveau.	Einführung
Aggregationsstufe	Jede Gliederung besitzt mehrere Aggregationsstufen, durch die Tabellenwerte mit Zwischen- und Randsummen festgelegt werden. Beim Regionalschlüssel z.B. gibt es die Aggregationsstufen *Gemeinde, Kreis, Regierungsbezirk, Land* und Bund (*Insgesamt).* Mit *Insgesamt* wird i.a. die jeweils höchste Aggregationsstufe bezeichnet. Eine Gliederung besitzt mindestens zwei Aggregationsstufen - die *Tabellenwerte* zu den Gliederungspositionen auf dem untersten Aggregationsniveau und die *Insgesamt-Summe*.	Einführung
Aggregations(stufen)index	Anzahl der Gliederungsstufen als Index, dabei ist die unterste die erste Gliederungsstufe und erhält den Index 1.	1.2
Aggregationsstufennummer	Anzahl der Gliederungsstufen, die unterste ist die erste Gliederungsstufe	1.2
aufgestockte Tabelle	Durch Einfügen von strukturellen Tabellenfeldern (Dummies) und Summen ergänzte Tabelle, in der nach der Ergänzung keine Zwischensummen mehr auftreten.	6.2
Berichtender	Meldeeinheit, z.B. natürliche oder juristische Person, die einen zu einem Aggregat beitragenden Wert gemeldet hat.	Zur Motivation
diametrale Werte	Zwei Tabellenwerte sind zueinander diametral, wenn sich ihre Gliederungsmerkmale bezüglich jeder Gliederung voneinander unterscheiden.	2.1.1
Differenzenverfahren	Verhindert die Rückrechnung gesperrter Werte durch Differenzbildung mit jeder Summe und der noch offenen Werten, die zu dieser Summe beitragen.	1.3
Dominanz	Dominanz liegt vor, wenn als bekannt vorauszusetzen ist, dass ein Wert oder die Wertesumme einer vorgegebenen Anzahl der größten Werte eines Aggregats einen vorgegebenen Anteil dieses Ag-	0.2.1 und 4.2.2

	ten Werte eines Aggregats einen vorgegebenen Anteil dieses Aggregats übersteigt. Durch das betreffende Aggregat ist dann der größte Einzelwert bzw. die Summe der größten Einzelwerte zu genau bestimmt, so dass dieses Aggregat nicht veröffentlicht werden kann. Dominanz hat auch für die sekundäre Geheimhaltung Bedeutung; weil man offenbar sehr große Werte nicht durch Sperren sehr kleiner schützen kann.	
Doppelquadersicherung	Sind bei der Sicherung des Pivots Einzelangaben als Quaderelemente zugelassen, Sicherung durch 2 Quader, damit kein Einzelmelder durch Eintrag seines Wertes den zu sichernden Wert berechnen kann.	2.1.1
Dummy	Bei Vervollständigung einer durch Zwischensummen unterteilten Tabelle in eine zwischensummenfreie Tabelle einzufügende zusätzliche, nicht zu veröffentlichende Tabellenwerte.	6.2.2
Durchstechen	Übernahme des Sperrmusters einer Tabelle (führendes Merkmal) auf andere nach den selben Gliederungen gegliederte Tabellen mit vergleichbarer Wertestruktur.	6.1.2
Einzelangabe	Tabellenwert, der von nur einem Berichtenden stammt	2.1.2
Einzeltabelle	Eine Einzeltabelle ist eine Menge von Aggregaten, die nach einer Kombination von mehreren hierarchischen Gliederungen gegliedert ist. Die Menge der Aggregate ist dabei vollständig, d.h. zu jeder Kombination von Schlüsseln der beteiligten Gliederungen gibt es genau ein Aggregat. Die Einzeltabelle heißt n-dimensional, wenn daran n Gliederungen beteiligt sind.	6.1
Einzelwert	Tabellenwert, der von nur einem Berichtenden stammt	0.2
Einzelwertsicherung	Einzelwertsicherung – wie z. B. beim Quaderverfahren - bezeichnet den Sicherungsprozess zum Schutze jedes einzelnen geheimen Tabellenwertes für sich allein, d.h. jeder einzelnen Einzelangabe und jedes anderen einzelnen geheimen Tabellenwertes, der von mehr als einem Berichtenden stammt. Im Gegensatz dazu kann auch – anders als beim Quaderverfahren - eine gleichzeitige (gemeinsame) Sicherung mehrerer geheimer Werte gefordert werden, so dass man nicht mehr von Einzelwertsicherung spricht. Dennoch sind bei Einzelwertsicherung wie z.B. beim Quaderverfahren gleichzeitige Sicherungen mehrerer geheimer Werte nicht ausgeschlossen, ja sogar erwünscht aber eben nicht zwingend Ziel des Prozesses einer Einzelwertsicherung.	2.1
Elementarindex	Der Elementarindex bezeichnet einen Index in einer zur vollständigen Tabelle aufgestockten (ursprünglich hierarchisch gegliederten) Tabelle, der nur einer Summierung entspricht. So ist jeder Summenbildung in der gegebenen Tabelle genau ein Elementarindex in der aufgestockten Tabelle zugeordnet, z.B. in der regionalen Gliederung einer den Gemeinden, einer den Kreisen und kreisfreien Städten sowie einer den Regierungsbezirken.	6.2
erlaubter Dummy	Dummy, der als Quaderwert eines zu sichernden Pivots genutzt werden kann.	6.2
Fallzahl	Eine Fallzahl ist ein Datenwert, der sich durch Zählung (+1) der Einzelfälle zu einem definierten Gliederungskontext aus dem Einzelmaterial ergibt. Summen von Fallzahlen sind wiederum Fallzah-	0.1

	len.	
Fallzahl-Additivität	Fallzahl-Additivität liegt vor, wenn in der betreffenden Gliederung sowohl die Tabellenwerte als auch ihre Fallzahlen addiert werden; sie ist nicht gegeben in Gliederungen, bei denen die Werte einzelner Berichtender gegliedert sind.	2.1.2.2
fiktive vollständige Tabelle	Eine „gedachte“, nicht real vorhandene, nicht im Hauptspeicher oder auf anderen Datenträgern gespeicherte vollständige Tabelle, die aber durch eine Verknüpfung ihrer Elementarindizes mit den Nutzerindizes auf die gegebene hierarchisch gegliederte Tabelle abgebildet werden kann.	7.
führendes Merkmal	Bei mehr als einem Tabellenwert pro Tabellenfeld kann die Geheimhaltung bezüglich eines ausgewählten Wertes vorgenommen und alle Sperrvermerke auf die übrigen Werte übertragen werden. Dieser ausgewählte Wert wird als führendes Merkmal bezeichnet.	6.1.2
ganz im Inneren einer Tabelle liegende Werte	Ein Tabellenwert liegt ganz im Inneren einer aufgestockten Tabelle, wenn er bezüglich jedes Elementarindexes 2 Nachbarwerte hat.	7.3
gerade indizierter Wert	Ein Quaderwert ist gerade indiziert, wenn sein Fehler ε positiv angesetzt wird (siehe Definition!).	3.1.2
Gewichtung (externe, interne)	Verändern des durch Sperren von Werten verursachten Informationsverlusts durch Multiplikation mit (von „außen“ frei vorgebbaren = externen oder durch Parameter zu steuernden = internen) Gewichtswerten	5.2.1 und 5.3
Gliederung	Gliederungen (im Gegensatz zu Wertmerkmalen) beinhalten sachliche Klassifizierungen wie z.B. die Wirtschaftszweigsystematik, Größenklassengliederungen, Altersgliederungen.	Einführung
Gliederungsmerkmal	Merkmal einer Kategorie der betreffenden Gliederung wie z.B. in der regionalen Gliederung eine spezielle Gemeinde oder ein spezieller Regierungsbezirk	Einführung
hierarchische Gliederung	Hierarchisch heißt eine Gliederung, deren feinste Kategorien zu zusammenfassenden gröberen Kategorien und die wiederum zu noch umfassenderen Kategorien zusammengefasst werden usw. bis schließlich alles unter einem Gesamtbegriff steht. Z.B. die regionale Gliederung in NRW ist eine hierarchische Gliederung. Darin werden die Gemeinden als feinste Kategorien zu den Kreisen, einer Vergröberung der Gemeinden, die Kreise (und kreisfreien Städte) zu den Regierungsbezirken, einer noch höheren Verdichtungsstufe, zusammengefasst, die dann schließlich unter dem Gesamtbegriff „Land“ stehen.	6.2
hinreichender Schutz	Die Sicherung eines geheimen Wertes ist hinreichend, falls dieser Wert nicht genauer als durch sein Schutzintervall festgelegt bestimmt werden kann.	3.1.2
im Inneren einer Teiltabelle	Ein Tabellenwert liegt im Inneren einer Teiltabelle, wenn er nicht in einer der diese Teiltabelle abgrenzenden Summen liegt. Ein Quader liegt im Innern einer Teiltabelle, wenn alle seine Elemente im Inneren der Teiltabelle liegen.	7.1.2
im Inneren einer (Unter-) Tabelle	Ein Tabellenwert liegt im Inneren einer (Unter-)Tabelle, wenn er nicht in einer ihrer Randsummen liegt. Ein Quader liegt im Innern einer (Unter-) Tabelle, wenn alle seine Elemente im Inneren der	1.2.1

	Untertabelle liegen.	
instantane Gewichtung	Gewichtung mit dem Abstand der Quaderwerte von ihrem Pivot im Augenblick der Quaderauswahl. Sie wird durch vom Tabellenschützer vorgebbare Parameter gesteuert und erfolgt unabhängig von der Vorgabe anderer Gewichte.	5.3.3
interne Gewichtung	siehe Gewichtung und programminterne Justierung	5.2.1
Intervallschutz	Intervallschutz setzt durch die Geheimhaltung Grenzen für ein Schutzintervall jedes geheimen Tabellenwertes, das ein Tabellennutzer nicht weiter einengen können soll.	3.
Justierung (intern, extern)	die gezielte Beeinflussung des Sperrmusters durch Gewichtung (siehe Gewichtung)	5.
Karree-Sicherung	In 2-dimensionalen Tabellen wird jedem geheimen Wert ein Karree mit lauter gesperrten Werten zugeordnet.	1.1
Mächtigkeit eines n-dimensionalen Quaders	Die Mächtigkeit eines n-dimensionalen Quaders ist die Anzahl seiner Quaderelemente. Sie beträgt 2^n.	2.2.1.1
Merkmal	Eine in einem Datenbestand oder in einer Tabelle niedergelegte Eigenschaft. Dabei werden nominal- oder ordinalskalierte Eigenschaften als qualitative Merkmale bezeichnet und meist als Gliederungsmerkmale genutzt, während intervallskalierte Werte oder Werte in einer Verhältnisskala als sogenannte quantitative Merkmale in Wertetabellen z.B. als Aggregate ausgewiesen werden.	Einführung
n-dimensionaler diskreter Raum	Die Klassifizierung der Daten einer Tabelle nach n Gliederungen liefert eine diskrete Gesamtheit von n-Tupeln zur Positionierung jedes Tabellenfeldes, die als kartesisches Produkt der Gliederungen angesehen und womit ein n-dimensionaler diskreter Raum beschrieben werden kann. Konkretisiert wird diese Raumvorstellung durch die instantane Gewichtung mit Hilfe eines Tabellenfeld-Abstandes.	2.1
n-dimensionale Tabelle	Eine n-dimensionale Tabelle ist eine Gesamtheit von Daten die nach n Gliederungen klassifiziert ist.	2.1
notwendiger Schutz	Unter notwendigem Schutz wird hier eine Maßnahme verstanden, die eine zu genaue Einschätzung einer Eigenschaft erschwert aber nicht verhindert, wie z.B. das Differenzenverfahren zur Vermeidung der Rückrechnung von Tabellenwerten durch Differenzbildung aus noch offenen Werten und Wertesummen.	Einführung
Nullwerte	1. Aggregate mit von Null verschiedenen Merkmalsträgeranzahlen 2. fiktive Werte in der Allgemeinheit nicht als leer bekannt vorauszusetzenden Tabellenfeldern, die als Sicherungspartner geheimer Werte verwendet werden können	3.1
Offener Wert	Aggregat, welches z.Zt. der Betrachtung nach bis zu diesem Zeitpunkt durchgeführter primärer und sekundärer Geheimhaltungs	5.3
	prüfung noch nicht gesperrt ist. Am Ende der Tabellensicherung	

	können dann alle noch nicht gesperrten Aggregate, d.h. die dann noch offenen Werte veröffentlicht werden.	
Paarigkeit von Einzelangaben	Einzelangaben sind paarig, wenn bezüglich einer Gliederung die Summe Einzelangabe ist und nur eine Einzelangabe (im Inneren) zu dieser Summe beiträgt. Bei Fallzahlen, die wie die Tabellenwerte aufsummiert werden (Fallzahl-Additivität), sind Summen-Einzelangaben immer paarig zu ihren Einzelangaben im Inneren.	2.1.2.2
partielle Aufstockung	Zur Sicherung eines Pivots wird aus der gegebenen Gesamttabelle temporär eine Teiltabelle so ausgewählt, dass für das Pivot zumindest ein vollständiger Sicherungsquader im Inneren der Teiltabelle existiert. Dieser Tabellenteil wird zur vollständigen Teiltabelle umstrukturiert, d.h. partiell aufgestockt.	6.2.2.3
Pivot(-wert, -element)	„Dreh- und Angelpunkt“, der während des Geheimhaltungsprozesses momentan zu sichernde geheime Wert.	2.1
Positionsindex	Ein Index für jede Gliederung zur Festlegung der geometrischen Position einer jeden Untertabelle. Jede n-dimensionale Untertabelle wird durch n Positions- und n Aggregationsindizes festgelegt.	1.2.3
positive Tabelle	Tabelle mit nicht negativen Tabellenwerten	Einführung
Primäre Geheimhaltung	Unter dem Begriff "Primäre Geheimhaltung" werden Maßnahmen zur Geheimhaltung verstanden, welche die zu genaue Schätzung von Einzelangaben auf der Grundlage des zugehörigen Aggregates verhindern.	Zur Motivation
Primärsperrungen	Bezeichnet die Menge der primär gesperrten Aggregate in einer Tabelle oder Aggregatdatei.	1.2
programminterne Justierung (durch interne Gewichtung)	bezeichnet die gezielte Beeinflussung der Verteilung von Sekundärsperrungen durch spezielle im EDV-Programm eingebaute Funktionen, die zum Teil durch Steuerparameter angesprochen werden können (z.B. Randschranken) oder die fest in das Programm integriert sind (z.B. Bevorzugung von geheimen vor offenen Werten als Sperrkandidaten durch negative Bewertung in der Quadersumme). (siehe auch Gewichtung)	5.2
Quader (n-dimensionaler)	Gesamtheit von Tabellenwerten, die durch zwei zueinander diametrale Tabellenwerte festgelegt ist. In zweidimensionalen Tabellen ist dies die Gesamtheit der Eckpunkte eines Karrees, in dreidimensionalen Tabellen die Gesamtheit der Eckpunkte eines geometrisch anschaulichen dreidimensionalen Quaders – daher auch die Bezeichnung n-dimensionaler Quader, wo n = 1; 2; 3; 4; 5;	1.3
Quaderaußenwerte	Als Quaderaußenwerte werden alle Tabellenwerte bezeichnet, die nicht der Gesamtheit geheimer Werte zum Schutz des betrachteten Pivots (Schutzgesamtheit) angehören, wobei die Schutzgesamtheit nicht notwendig die eines Quaders oder Doppelquaders sein muss.	7.3.5
Quaderdeformation	Zur Vermeidung von Summensperrungen können in aufgestockten Tabellen Quaderteilgesamtheiten gegeneinander so verschoben werden, dass dabei als Quaderelemente unzulässige Werte gemieden werden. Diese Verschiebungen sind so vorzunehmen, dass die damit verbundenen Summenänderungen nur die nicht zu veröffentlichenden Sternchensummen betreffen.	7.3
Quaderspannweite	Die jedem Quader mit lauter geheimem Werten zuzuordnende Schutzintervalllänge.	3.

Quaderverfahren	Verfahren der sekundären Geheimhaltung, welches momentan durch die Programme GHMITER und QUIT realisiert wird. Das Quaderverfahren sucht für jeden zu schützenden Tabellenwert einen Quader von Tabellenwerten so aus, dass der zu schützende Pivotwert nicht genauer bestimmt werden kann, als es ein vorgegebenes Schutzintervall erlaubt, und sperrt die noch offenen Werte. Bei der Quaderauswahl wird in erster Linie die Anzahl zu sperrender Werte minimiert und dann erst eine möglichst kleine Summe gesperrter Werte angestrebt.	Zur Motivation
Randschranken	Um Randsummen bei den Sekundärsperrungen nach Möglichkeit zu meiden oder besonders zu bevorzugen, können positive oder negative Werte zu den betreffenden Randsummen ausgewählter Dimensionen addiert werden. Dies erfolgt durch setzen von Randschranken, die in GHMITER und in QUIT durch Steuerparameter vorgegeben werden können.	5.2.1
Randsumme	Randsummen sind Aggregate, die bzgl. mindestens eines Gliederungsmerkmals die maximale Aggregationsstufe haben.	Einführung
range	Kurzbezeichnung für den in der Statistik definierten Begriff der Spannweite.	Einführung
relative Mindestspannweite	Relative Spannweite, die bei der Quaderauswahl durch die Quaderspannweite zu überschreiten ist.	3.1.2.3
relative Spannweite	Spannweite bezogen auf den zugehörigen Wert.	3.1.2
Sachschlüssel	Merkmalsausprägungen einer sachtheoretischen Gliederung in der betreffenden Tabelle	Einführung
Schätzfehler	1. Allgemein: Betrag der oberen bzw. unteren Abweichung des (vom Tabellennutzer) geschätzten Wertes von seinem tatsächlichen Wert. 2. Als zu unterstellende Nutzerinformation: Größtmöglicher Betrag der oberen bzw. unteren Abweichung eines vom Tabellennutzer geschätzten Wertes von seinem tatsächlichen Wert.	Einführung
Schätzintervall	(Als Vorwissen beim Tabellennutzer zu unterstellender) Wertebereich eines (vom Tabellennutzer) zu schätzenden Wertes, d.h. Wertebereich zwischen dem größtmöglichen und dem kleinstmöglichen Schätzwert, den man dem Tabellennutzer noch zutraut.	Einführung
Schloss	Gliederungsmerkmal, das als Steuerungsinstrument vom Geheimhaltungsprogramm benutzt wird, um die Gliederungsmerkmale nach dem Schlüssel-Schloss-Prinzip der Tabellenstruktur zuzuordnen.	Einführung
Schlossdatei	Gesamtheit der Schlösser: umfasst zu jeder der n Gliederungen mindestens alle ihre Gliederungsausprägungen und ist innerhalb jeder Gliederung nach der Höhe der Aggregationsstufen sortiert.	Einführung
Schlüssel einer Gliederung (fachlich, technisch)	Für die Kennzeichnung der einzelnen Ausprägungen einer Gliederung sind numerische, hierarchie-konform aufgebaute Schlüssel erforderlich. Solche Schlüssel genügen folgenden Regeln: - Als Ausprägung von Teilschlüsseln kommen keine Nullen vor.	Einführung

	- Verkürzte Schlüssel für höhere Stufen werden nach rechts durch Nullen auf die einheitliche Länge aufgefüllt. - Für die Totalsumme wird links eine Hierarchiestufe mit der Ausprägung 1 vorangestellt; sie kann technisch aber auch durch Nullen in allen Schlüsselstellen repräsentiert werden. *Fachliche Schlüssel* können daher nur verwendet werden, wenn zu der Gliederung eine Systematik existiert, die o.g. Regeln erfüllt. Ist dies nicht der Fall, so muss ein entsprechend aufgebauter *technischer Schlüssel* definiert und die Daten umgeschlüsselt werden.	
Schutzgesamtheit	Die Gesamtheit aller für eine hinreichende Sicherung eines geheimen Wertes erforderlichen gesperrten Tabellenwerte wird als Schutzgesamtheit (dieses Wertes) bezeichnet. Beim Quaderverfahren ist dies die Gesamtheit der Elemente eines (deformierten) Quaders oder Doppelquaders, der für die Sicherung eines geheimen Wertes, das Element des Quaders oder Doppelquaders ist, hinreichend ist.	2.1
Schutzintervall (eines geheimen Wertes)	Intervall das einen geheimen Tabellenwert überdeckt, nachdem eine Sicherung der Tabelle mit einem Verfahren zur Geheimhaltung vorgenommen wurde.	3.1
Segment (einer hierarchischen Gliederung)	Gliederungsmerkmale, die als Indizes zu einer Summenbildung gehören.	6.2
Sekundäre Geheimhaltung	Unter dem Begriff "Sekundäre Geheimhaltung" werden zusätzliche Maßnahmen zur Sicherstellung der Geheimhaltung verstanden, welche die näherungsweise Offenlegung von primär geheimen Werten aus dem Zusammenhang der veröffentlichten Aggregate heraus und unter Hinzuziehung von Vorwissen verhindern.	Zur Motivation
Sekundärsperrungen	Bezeichnet die Menge der sekundär gesperrten Aggregate in einer Tabelle oder Aggregatdatei.	Einführung
sperrbare Null	Wert einer Tabelle mit Ausprägung Null, der als Quaderelement in Frage kommt, weil er als hinreichend ungenau schätzbar vorausgesetzt werden kann.	7.3
Sternchensummen	Bei Vervollständigung einer durch Zwischensummen unterteilten Tabelle in eine zwischensummenfreie Tabelle einzufügende zusätzliche, nicht zu veröffentlichende Summen.	6.2.2
Statistik	Eine Statistik ist die Gesamtheit der (z.B. durch Meldebögen) erhobenen Daten, die nach einer Teilgesamtheit dieser Daten, der Gesamtheit aller r Gliederungsmerkmale, gegliedert ist, d.h. die in einem von diesen r Gliederungsmerkmalen aufgespannten Raum, dem kartesischen Produkt dieser Gliederungsmerkmale, eingeordnet werden können.	6.1.2
Strang primär geheimer Werte	Gesamtheit primär geheimer Werte, die zu einem primär geheimen Wert höherer Aggregationsstufe beitragen.	7.2
strukturelle Nullen	Wert einer Tabelle mit Ausprägung Null, der nicht als Quaderelement in Frage kommt	5.1.2.4

	ment in Frage kommt.	
Summenwert	Summenwerte (im Gegensatz zu Fallzahlen) sind Aggregate, die sich durch die Addition von Einzelwerten aus dem Einzelmaterial oder von Summenwerten ergeben.	Zur Motivation
Tabellenfeld	Ein Tabellenfeld ist eine Position in einer n-dimensionalen Tabelle, die durch die Merkmalsausprägungen der Gliederungen fixiert ist (vgl. Zelle).	Einführung
Tabelleninneres	Das Tabelleninnere bezeichnet die Gesamtheit der Tabellenwerte ohne die Werte der Randsummen bezüglich jeder Gliederung.	3.1.2.1
Tabellenrand	Der Tabellenrand bezeichnet die Gesamtheit der Randsummen, die Gesamtheit der höchsten Aggregate bezüglich jeder Gliederung.	2.1.2.2
Teiltabelle	Eine Teilgesamtheit einer hierarchisch gegliederten Tabelle, die für jede Gliederung eine Randsumme hat, daneben können – im Unterschied zu den Untertabellen - auch Zwischensummen vorkommen.	6.2
überlappende Quader	Quader überlappen einander, wenn sie Quaderwerte gemeinsam haben.	3.2.3.3
überlappende (Einzel-) Tabellen	Tabellen überlappen einander, wenn sie Aggregate bzw. Datensätze gemeinsam haben. Die Menge dieser gemeinsamen Datensätze wird als Überlappungsbereich bezeichnet.	1.3
Umbuchung (von Fallzahlen)	Durch Übertragung von Tabelleneintragungen bzw. von Teilen davon in andere Tabellenfelder werden nicht zu veröffentlichende Tabellenwerte für die Veröffentlichung freigemacht. Insbesondere wird durch Übertragung von Fallzahlen erreicht, dass in jedem Tabellenfeld eine vorgegebene Mindestfallzahl nicht unterschritten wird, damit diese Werte nicht wegen zu geringer Fallzahlen gesperrt werden müssen.	4.1
ungerade indizierter Wert	Ein Quaderwert ist ungerade indiziert, wenn sein Fehler ε negativ angesetzt wird (siehe Definition!).	3.1.2
Untertabelle (n-dimensionale)	Untertabellen sind Teilgesamtheiten der n-dimensionalen Gesamttabelle, die bezüglich jeder der n Gliederungen nur eine Rand- und keine Zwischensummen aufweisen.	1.2.1
Untertabellenabgleich	Nach der Bearbeitung jeder einzelnen Untertabelle werden die im laufenden Bearbeitungsschritt eingetragenen Sperrvermerke (samt Schutzintervallen) in die Gesamttabelle übertragen und so dafür gesorgt, dass alle mehreren Untertabellen gemeinsamen Werte den selben Geheimhaltungsstatus haben.	1.2.3
Untertabellenhierarchie	In einer hierarchisch gegliederten Tabelle gibt es Untertabellen zu allen Aggregationsstufen der n Gliederungen. Hinsichtlich ihrer Verdichtung, der Höhe der Aggregationsstufen ihrer Gliederungen, haben diese Untertabellen eine hierarchische Struktur, sie bilden eine Untertabellenhierarchie.	1.2.1
Veröffentlichungsraum	bezeichnet den p-dimensionalen Raum, der von allen p Gliederungen aller Veröffentlichungstabellen aufgespannt wird.	6.1.2
Veröffentlichungstabelle	ist eine (i.d.R. nur nach wenigen Merkmalen gegliederte) Einzeltabelle.	6.1.2
verbotener Dummy	Dummy, der als Quaderwert eines zu sichernden Pivots nicht genutzt werden kann	6.2

	nutzt werden kann.	
Vollständiger Quader	Ein Quader ist vollständig, wenn seine Elemente ausschließlich in einer einzigen zwischensummenfreien (Unter-)Tabelle vorkommen. D.h. der Quader ist vollständig, wenn keines seiner Elemente in einer anderen Tabelle gesichert wird.	3.2.3.3
Vollständige Tabelle	Eine Statistiktabelle heißt vollständig, wenn die Addition von Tabellenwerten über jedes Gliederungskriterium (über jeden Index) immer zu genau einer Summe , der Randsumme, führt; die Gliederungskriterien, die zu keiner Zwischensumme Anlass geben, werden als elementare Gliederungen bezeichnet, ihre Indizes als Elementarindizes.	Einführung u. 6.2.2
Vorinformation (Vorwissen)	Vorinformation oder Vorwissen ist das Wissen, das der Tabellennutzer ohne Kenntnis der Tabelle haben kann und das man folglich bei ihm unterstellen muss.	1.2.1 und 3.2
Vorlauftabelle	Als Vorlauftabellen im weitesten Sinne werden hier Tabellen bezeichnet, die bereits veröffentlicht sind und mit der zu sichernden Tabelle in Beziehung stehen, so dass wechselseitige Rückschlüsse auf geheime Werte zu befürchten sind. Als Beispiel sind zeitlich aufeinanderfolgende Veröffentlichungstabellen mit einheitlicher Gliederungsstruktur, sogenannte Zeitreihen (-Tabellen) zu nennen. Geheime Werte in der aktuellen Tabelle können z.B. durch Übernahme der entsprechenden offenen Werte aus der Vorperiode oft recht genau geschätzt werden.	3.2.3.1
Wertartschlüssel	Der Wertartschlüssel belegt ein Feld im Datensatz der Tabelle. Er speichert die Information, ob der Wert des Datensatzes veröffentlicht werden darf oder ob er geheimgehalten werden muss.	Einführung
Werte einer Tabelle	Rationale Zahlen mit einer endlichen Stellenzahl, die aus den entsprechenden Angaben der Berichtenden eines jeden Tabellenfeldes durch Addition entstehen, sie werden daher auch als Aggregate bezeichnet.	Zur Motivation
Werteklassierung	Durch logarithmische Klassierung aller Werte der Tabelle und mehrfache Aneinanderreihung dieser abgeschlossenen Gesamtheit entlang des Zahlenstrahls entsteht eine hierarchische Klassenstaffelung, die in allen Versionen von GHMITER und QUIT als Träger qualitativer Wert-Merkmale wie primär geheim, Einzelwert, sekundär geheim, ... genutzt wird.	Einführung u. 5.2
Werteverfälschung (Perturbation)	Verfahren zum Schutze nicht zu veröffentlichender Werte, bei dem der Wert soweit verändert wird, dass danach seine Veröffentlichung nicht gegen die Geheimhaltungsrichtlinien verstößt.	4.1
Zeitreihentabellen	Zeitreihentabellen sind zeitlich aufeinanderfolgende Veröffentlichungstabellen mit einheitlicher Gliederungsstruktur. Diese Bezeichnung wird hier auch synonym für Tabellen verwendet, die mit der zu sichernden Tabelle in Beziehung stehen, so dass wechselseitige Rückschlüsse auf geheime Werte zu befürchten sind (siehe auch Vorlauftabellen).	3.2.3.1
Zelle einer n-dimensionalen Tabelle	Eine Zelle ist eine durch konkrete Ausprägungen der n Gliederungen bestimmte Position in der n-dimensionalen Tabelle.	1.3
Zwischensumme	Eine Zwischensumme ist die Summe von Tabellenwerten bezüglich einer Gliederung, die selbst mit anderen Zwischensummen zu einer	1.2

Zwischensumme höherer Aggregation oder zur Randsumme beiträgt.

C Kern des Quaderverfahrens[16]

Gegeben sei eine n-dimensionale Tabelle T mit Randsummen bezüglich jedes der n Gliederungen aber ohne Zwischensummen. Die Ausprägungen jeder Gliederung seien durchnummeriert, die Tabellenwerte mit diesen Nummern indiziert.

Definition 1:

Zwei Tabellenwerte $T_a, T_b \in T$ heißen zueinander diametral, symbolisch $T_a \# T_b$, wenn für ihre Indizes a $= (a_1,a_2,...,a_n)$, $b = (b_1,b_2,...,b_n)$ gilt

$$a_i \neq b_i, \quad \text{für } i = 1,2, \ldots, n .$$

Definition 2:

$Q(T_a, T_b) = \{ T_q : T_a, T_b, T_q \in T; T_a \# T_b; q_i(k) = a_i + (b_i - a_i)* B_{ik}\}$ heißt n-dimensionaler Quader, wobei B_{ik} = i-te Binärstelle von k, k = 0,1,2,...,2^n-1; k ist die Nummer des Quaderwertes $T_q = T_q(k) \in Q(T_a, T_b)$.

Definition 3:

$T_q(k) \in Q(T_a, T_b)$ heißt bezüglich T_b gerade indiziert, wenn $\sum_i B_{ik} + \sum_j A_j$ gerade ist, anderenfalls heißt $T_q(k)$ ungerade indiziert, k = 0, 1, ..., 2^n-1. Dabei bezeichnet A_j die Aggregationsstufe von $T_q(k)$ bezüglich der j-ten Gliederung, j = 1,2,...,n; die Aggregationsstufen sind mit 1 beginnend in Einserschritten fortlaufend durchnummeriert (T_a = Pivot).

Die linearen Gleichungen zur Berechnung gesperrter Werte $X,Y,X',Y' \in Q(T_a, T_b)$, wo X,Y bezüglich T_b gerade, X′,Y′ bezüglich T_b ungerade indiziert sind, haben die Gestalt:

$$X + X' = \Sigma , \quad X - Y = \Sigma , \quad X' - Y' = \Sigma \qquad (1),$$

wobei Σ die Quadersumme bezeichnet: Σ ist die Randsumme abzüglich aller nicht zu $Q(T_a, T_b)$ gehörenden Tabellenwerte bezüglich der Summations-Gliederung. Tragen keine Randsummenwerte zu $Q(T_a, T_b)$ bei, gilt nur die erste Gleichung von (1).

Falls alle Tabellenwerte $X, X' \in Q(T_a, T_b)$ gesperrt sind, kann der externe Tabellennutzer nur Schätzwerte für X,X′ angeben:

$$\hat{X} = X + \varepsilon ; \quad \hat{X}' = X' - \varepsilon \qquad (2);$$

mit nur einem unbekannten Parameter ε ∀ $X,X' \in Q(T_a, T_b)$, der – falls kein Vorwissen unterstellt werden muss – beliebige Werte annehmen kann.

---

[16] Diese Zusammenfassung entstammt dem Anhang zum Aufsatz des Autors „Sicherung persönlicher Angaben in Tabellendaten“, erschienen in Statistische Analysen und Studien NRW, Ausgabe 1/ 2002.

Bei zu unterstellendem Vorwissen kennt der Tabellennutzer Schätzintervalle mit Intervallgrenzen T_{oben}, T_{unten} $\forall$ T_t $\in$ T, so dass auch für die Schätzwerte der Gleichungen (2) gilt,

$$\mathbf{X_{unten} \leq X + \varepsilon \leq X_{oben}\,, \quad X'_{unten} \leq X' - \varepsilon \leq X'_{oben}} \qquad \textbf{(3).}$$

Dadurch wird der Schätzfehler ε eingegrenzt gemäß

$$\mathbf{-\varepsilon_{unten} \leq \varepsilon \leq \varepsilon_{oben}} \qquad \textbf{(4),}$$

wobei die Beträge der Intervallgrenzen gemäß (3) und (4), zusammenfassend dargestellt, die Gestalt haben:

$$\mathbf{\varepsilon_{oben} = min((X_{oben} - X), (X' - X'_{unten}));} \quad \mathbf{\varepsilon_{unten} = min((X'_{oben} - X'), (X - X_{unten}))} \qquad \textbf{(5).}$$

Jeder Quaderwert X bzw. X´ $\in$ Q(T_a, T_b) kann vom Tabellennutzer somit höchstens bis auf sein Schutzintervall genau eingegrenzt werden:

$$\mathbf{\hat{X} \in [\, X - \varepsilon_{unten}, X + \varepsilon_{oben}\,], \quad \hat{X}' \in [\, X' - \varepsilon_{oben}, X' + \varepsilon_{unten}\,]} \qquad \textbf{(6).}$$

Die Spannweite des Quaders Q(T_a, T_b), range, ist dem gemäß

$$\mathbf{range = \varepsilon_{unten} + \varepsilon_{oben}} \qquad \textbf{(7).}$$

Damit ist ein Quaderauswahlkriterium gegeben, das in den derzeit verfügbaren EDV-Programmen GHQUAR.44 und GHMITER.22 angewendet wird und das auch in QUIT die Quaderauswahl steuert: Es kommen nur solche Quader als Sicherungsquader in Betracht, deren Spannweite, range, bezogen auf den Betrag des zu schützenden Pivotwertes größer als eine vorgegebene relative Mindestspannweite ist. Bei Nullwerten und besonders kleinen Werten wird man einen Absolutwert als nicht zu unterschreitende Schranke für die Quaderauswahl vorgeben.

Literaturangaben

Günter Appel, S. Kinzel, D. Nölte, Statistisches Landesamt Berlin:

SAFE - A Generally Usable Program System for the Anonymization of Individual Data in Official Statistics, Internat. Semin. on Statist. Confid., Dublin 1992

Lawrence H. Cox, U.S. Bureau of the Census:

1) Suppression Methodology and Statistical Disclosure Control, Journal of the American Statistical Association, 1980, Vol. 75, No 370

2) Solving Confidentiality Protection Problems in Tabulations using Network Optimization: A Network Model for Cell Suppression in U.S. Economic Censuses, International Seminar on Statistical Confidentiality, Dublin 1992

Jacquelin Geurts, Netherlands Central Bureau of Statistics:

Heuristics for Cell Suppression in Tables, Department of Statistical Methods, P.O. Box 959, 2270 AZ Voorburg, The Netherlands 1992

Sarah Gießing, Statistisches Bundesamt:

„Statistische Geheimhaltung in Tabellen“, Forum der Bundesstatistik, Bd. 31/1999

Rüdiger D. Repsilber, Landesamt für Datenverarbeitung u. Statistik NRW:

1) EDV-Verfahren zur Wahrung der Geheimhaltung bei Tabellen mit bis zu sieben Ordnungskriterien, Statist. Rundschau Nordrhein-Westfalen, Landesamt für Datenverarbeitung u. Statistik Nordrhein-Westfalen 1991

2) Safeguarding Secrecy in Aggregative Data, Internat. Seminar on Statist. Confid., Dublin 1992

3) Preservation of Confidentiality in Aggregated Data, Internat. Seminar on Statist. Confid., Luxembourg 1994

4) Das Quaderverfahren, Forum der Bundesstatistik, Band 31/1999

5) Wahrung der Geheimhaltung sensibler Daten in mehrdimensionalen Tabellen mit dem Quaderverfahren, Statistische Analysen und Studien Nordrhein-Westfalen, Ausgabe 3/2000

Dale A. Robertson, Statistics Canada:

Automated Disclosure Control at Statistics Canada, International Seminar on Statistical Confidentiality, Luxembourg 1994

Laura V. Zayatz, U.S. Bureau of the Census:

Using Linear Programming Methodology for Disclosure Avoidance Purposes, Internat. Seminar on Statistical Confid., Dublin 1992