Statistische Analysen und Studien Nordrhein-Westfalen

Ausgabe 3/2000

Impressum

Herausgeber :
Landesamt für Datenverarbeitung und Statistik Nordrhein-Westfalen

Redaktion:
Jörg Mühlenhaupt, Hans Lohmann

Preis dieser Ausgabe: 6,50 DM

Erscheinungsfolge: unregelmäßig

Bestellungen nehmen entgegen:

das Landesamt für Datenverarbeitung und Statistik NRW,
Postfach 10 11 05,
40002 Düsseldorf,
Mauerstraße 51,
40476 Düsseldorf
Telefon: (02 11) 94 49-25 16/35 16
Telefax: (02 11) 44 20 06
Internet: http://www.lds.nrw.de
E-Mail: poststelle@lds.nrw.de

sowie der Buchhandel.

Pressestelle:
(02 11) 94 49-25 21/25 18

Zentraler Informationsdienst:
(02 11) 94 49-24 95/25 25

Bestell-Nr. Z 08 1 2000 53

Inhalt

editorial

Die Gewährleistung der statistischen Geheimhaltung, d. h. die Vermeidung der Offenlegung persönlicher Daten, ist eine fundamentale Aufgabe jeder Statistiken erhebenden und verbreitenden Institution, weil damit die für die Aussagefähigkeit der Daten unabdingbare Vertrauensbasis geschaffen und erhalten 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, so viel Offenlegung wie möglich und nur so viel Geheimhaltung wie unbedingt nötig vorzusehen. 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 weitere vertrauensbildende Maßnahme zur Förderung der Akzeptanz von Statistikerhebungen kommt der Offenlegung der angewendeten Verfahren zur Wahrung der Geheimhaltung eine große Bedeutung zu, insbesondere dann, wenn diese Verfahren ganz gezielt so entwickelt wurden, dass sie – zumindest im Prinzip – allgemein verständlich darstellbar sind. Das gilt auch für eine umfassende Darstellung der Weiterentwicklung von Geheimhaltungsverfahren, die dem veröffentlichenden Statistiker auf Grund von immer effizienter werdender so genannter Attackersoftware aufgezwungen wird. Der amtlichen Statistik ist daher die Entwicklung eines einfachen EDV-Verfahrens zur Wahrung der Geheimhaltung ein besonderes Anliegen, wobei sie nicht nur die Nutzer der hinsichtlich der statistischen Geheimhaltung gesicherten Tabellen im Blickfeld hat, sondern auch die die Statistiken vertreibenden Fachstatistiker, die mit Hilfe solcher EDV-Verfahren für die geforderte Datensicherheit zu sorgen haben. Der Verfasser des hier vorliegenden Textes stellt ein von ihm im LDS NRW entwickeltes Verfahren vor.

Gegenstand des im Folgenden diskutierten Geheimhaltungsverfahrens sind nach mehreren Gliederungskriterien gegliederte, oft mehrfach durch Zwischensummen unterteilte Statistiktabellen. In solchen fein gegliederten Tabellen treten oft viele Werte auf, die einzelnen Berichtenden zugeordnet werden können und die z. B. durch Sperren dieser Werte geheim gehalten werden müssen. Die Unterdrückung dieser sensiblen Werte in der Veröffentlichungstabelle 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. Als Vorwissen sind diejenigen Kenntnisse zu verstehen, die der Tabellennutzer noch vor der Veröffentlichung der Tabelle haben kann, mit dem er also die Tabellenwerte eingrenzen kann. Die Verhinderung der zu genauen Rückrechnung primär geheimer Werte aus noch offenen Tabellenwerten unter Zuhilfenahme von zu unterstellendem Vorwissen wird als sekundäre Geheimhaltung bezeichnet. Ein besonders einfaches Verfahren zur sekundären Geheimhaltung ist das Quaderverfahren, das das Thema dieser Arbeit ist. Das vorgestellte Verfahren wurde international mit Aufmerksamkeit wahrgenommen; das Statistische Amt der Europäischen Gemeinschaften (EUROSTAT) beabsichtigt, das Quaderverfahren einzusetzen und es den Ländern der EU als praktikables Verfahren zur Wahrung der Geheimhaltung in umfangreichen Statistiktabellen vorzuschlagen.

Jochen Kehlenbach

Präsident

Wahrung der Geheimhaltung sensibler Daten in mehrdimensionalen Tabellen mit dem Quaderverfahren

Rüdiger Dietz Repsilber

Im Folgenden wird 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 beschrieben. 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.

Grundlagen

Einführung

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 Nachweisungsfälle, den Wert des Nachweisungsmerkmals als Summe der Einzelmerkmalswerte und den Wertartschlüssel, d. h. die Kennzeichnung, ob der Wert geheim zu halten ist oder nicht, umfasst.

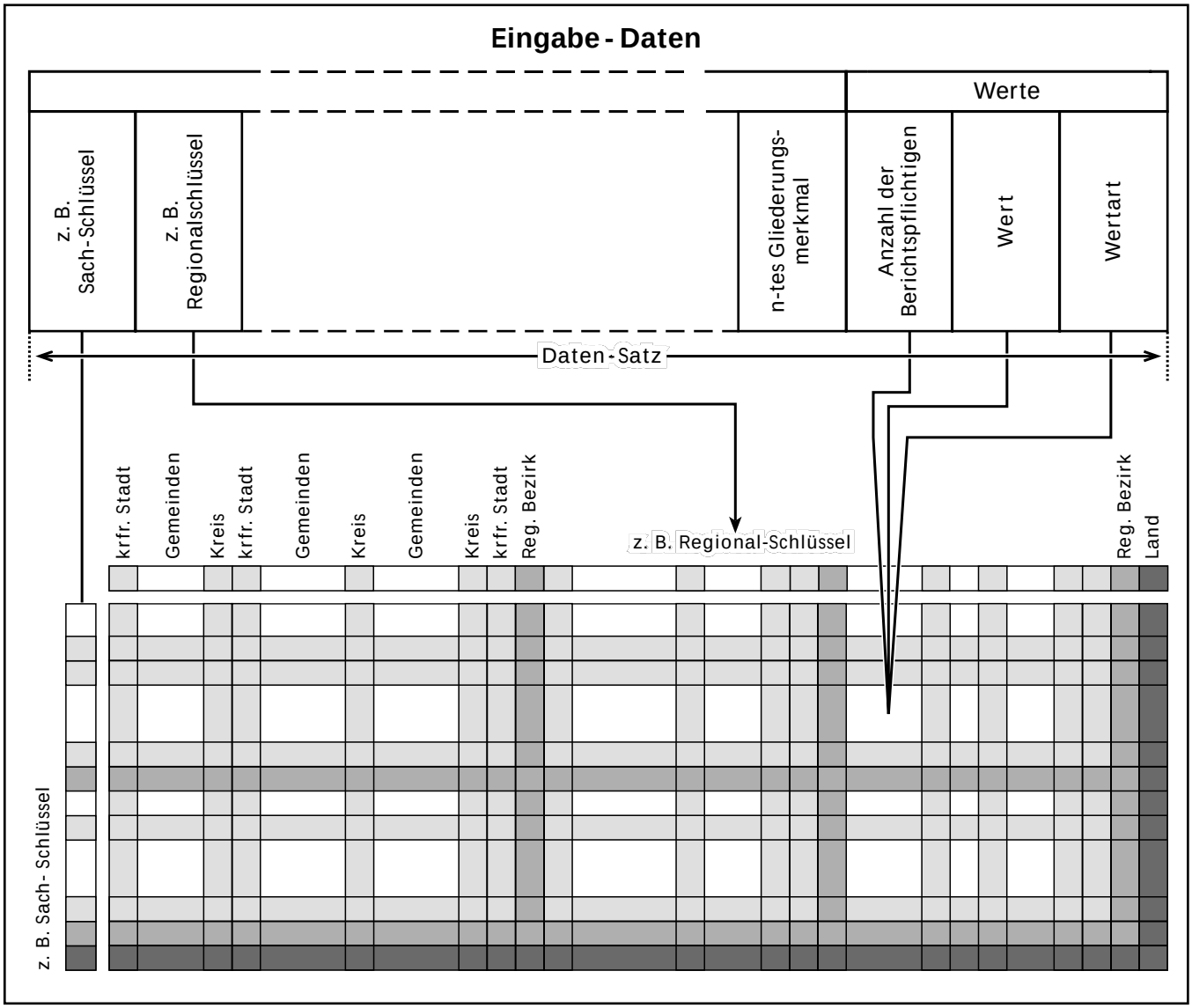
Die Abbildung „Eingabe-Daten“ zeigt in ihrem oberen Teil den Datensatz einer n-dimensionalen Tabelle. Im unteren Teil der Abbildung ist eine Tabelle für den Fall zweier Gliederungsmerkmale (z. B. ein Sach- und ein Regionalschlüssel) als zweidimensionales Zahlentableau in Gestalt eines Rechtecks dargestellt, 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 Schraffur zu ihrem jeweils darunterliegenden 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 „Eingabe-Daten“ durch eine regionale Gliederung ausgeführten Spaltengliederung erfolgt von links nach rechts, indem die Gemeinden zu ihren Kreisen, die Kreise und kreisfreien Städte zu ihren Regierungsbezirken und die Regierungsbezirke schließlich zum Land 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 „Eingabe-Daten“ 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 den Aufbau einer 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 denselben 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 Dimensi-



on 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 Dimension der aufgestockten Tabelle, die zur Unterscheidung von der ursprünglichen, durch die Anzahl der Gliederungskriterien bestimmten Dimension hier als Aggregatdimension bezeichnet werden soll. Bei mehrfach durch Zwischensummen unterteilten Tabellen kann erst die durch Dimensionsaufstockung umstrukturierte erweiterte Tabelle mit dem Quaderverfahren hinreichend geschützt werden.

Viele Ansätze zur Wahrung der Geheimhaltung bei möglichst geringem Informationsverlust sind – obwohl mathematisch als Optimierungsproblem exakt lösbar – 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). 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 Ge-

heimhaltung 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 Ab. 1.2). Umgekehrt werden, wenn die Anzahl der Sekundärsperrungen minimiert 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. f(X) = lnX eingeführt werden muss, um hier die gewünschten Prioritäten einzuhalten. Zwar benutzt auch das derzeit das Quaderverfahren realisierende EDV-Programm GHQUAR ebenfalls den Logarithmus der Werte an Stelle der Tabellenwerte selbst; dies dient aber ausschließlich einer auch die einzelnen Tabellenwerte noch genügend stark differenzierenden Werteklassierung, mit der die Information über die Wertattribute, primär geheim, Einzelangabe, sekundär geheim auf die Werte selbst übertragen wird, hat aber 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 temporär geeignet modifiziert werden, so dass sie auf Grund ihrer Größe bevorzugt zu Sekundärsperrungen herangezogen oder besonders gemieden werden. Auf die Justierungsmöglichkeiten der Verteilung der Sekundärsperrungen durch Gewichtung von Werten wird im fünften Abschnitt und am Schluss bei der Behandlung von Beispielen noch gesondert eingegangen.

1. Grundlegende Probleme 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 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.

1.2 Sekundäre Geheimhaltung mit 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 1.4). 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 gewährleistet, bietet sich bei 2-dimen-

Abb. 1.1

Summensperrungen

Kreis a = Anzahl b = Betrag		Gruppe A	Gruppe B	Gruppe C	Gruppe D	Σ
1	a	11	8	4	117	**140**
	b	7 760	240	57	4 154	**12 211**
2	a	3	●—2	33	67	⊙—105
	b	240	—187	184	1 782	—2 393
3	a	322	3	18	●—8	⊙—351
	b	1 723	316	115	—258	—2 412
4	a	116	87	21	4	**228**
	b	842	448	439	86	**1 815**
Regierungsbezirk	**a**	**452**	⊙—100	**76**	⊙—196	**824**
	b	**10 565**	—1 191	**795**	—6 280	**18 831**

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

sionalen Tabellen die Karree-Sicherung an, wobei jedem primär geheimen Wert – unabhängig von möglichen anderen geheimen Tabellenwerten – ein Karree mit lauter gesperrten Werten zugeordnet wird. 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.

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.

Eine nach diesen Kriterien gesicherte Tabelle zeigt die Abbildung 1.2.

Abb. 1.2

Minimale gesperrte Wertesumme

Kreis (a = Anzahl, b = Betrag)		Gruppe A	Gruppe B	Gruppe C	Gruppe D	Σ
1	a	11	⊙ ~~8~~	⊙ ~~4~~	117	**140**
	b	7 760	~~240~~	~~57~~	4 154	**12 211**
2	a	3	● ~~2~~	⊙ ~~33~~	67	**105**
	b	240	~~187~~	~~184~~	1 782	**2 393**
3	a	322	3	⊙ ~~18~~	● ~~8~~	**351**
	b	1 723	316	~~115~~	~~258~~	**2 412**
4	a	116	87	⊙ ~~21~~	⊙ ~~4~~	**228**
	b	842	448	~~439~~	~~86~~	**1 815**
Regierungsbezirk	**a**	**452**	**100**	**76**	**196**	**824**
	b	**10 565**	**1 191**	**795**	**6 280**	**18 831**

● = 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.

1.3 Sekundäre Geheimhaltung mit 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.

Die Abbildung 1.3 zeigt nun den gegenwärtig benutzten Ansatz: Der zur

Abb. 1.3

Minimale Anzahl Sekundärsperrungen

Kreis (a = Anzahl, b = Betrag)		Gruppe A	Gruppe B	Gruppe C	Gruppe D	Σ
1	a	11	8	4	117	**140**
	b	7 760	240	57	4 154	**12 211**
2	a	3	● ~~2~~	33	⊙ ~~67~~	**105**
	b	240	~~187~~	184	~~1 782~~	**2 393**
3	a	322	⊙ ~~3~~	18	● ~~8~~	**351**
	b	1 723	~~316~~	115	~~258~~	**2 412**
4	a	116	87	21	4	**228**
	b	842	448	439	86	**1 815**
Regierungsbezirk	**a**	**452**	**100**	**76**	**196**	**824**
	b	**10 565**	**1 191**	**795**	**6 280**	**18 831**

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

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 derselben 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).

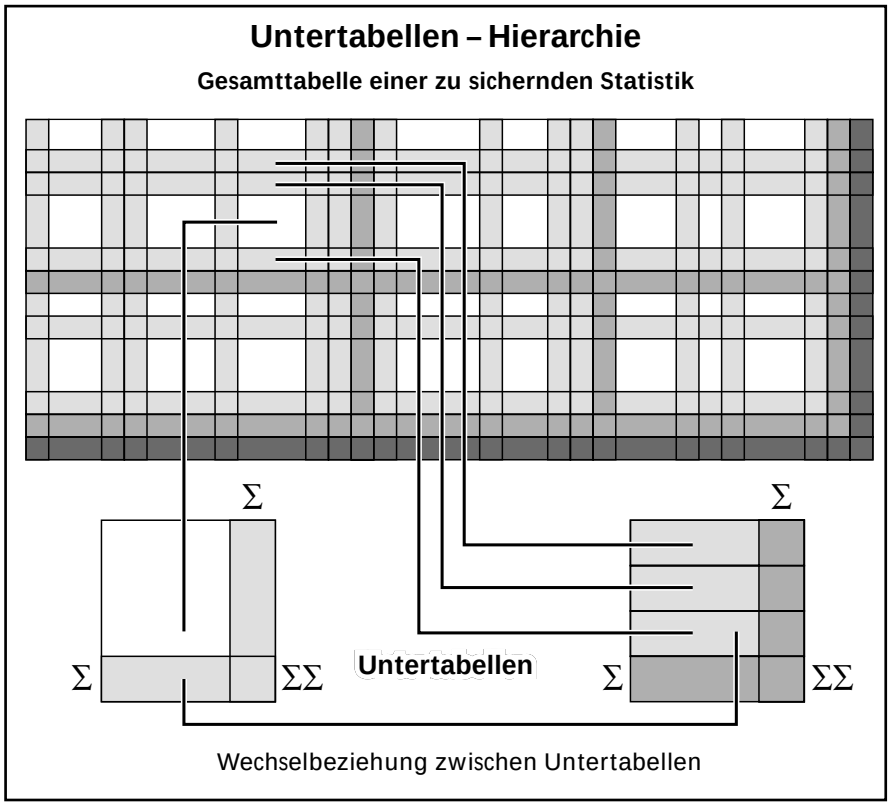
1.4 Sekundäre Geheimhaltung bei durch Zwischensummen untergliederten Tabellen

1.4.1 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, sie werden im Folgenden als Untertabellen bezeichnet. Im Falle der eingangs aufgeführten zweidimensionalen „Eingabe-Daten“ (unterer Teil der Abbildung Eingabe-Daten) 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



Wechselbeziehung zwischen Untertabellen

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.

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.

Die Sicherung jeder Untertabelle für sich alleine betrachtet, d. h. herausgelöst aus ihrer Untertabellenhierarchie, stellt eine aus der Sicht der Gesamttabelle im Allg. unzulässige Idealisierung dar, weil Sperrungen in Untertabellen höherer Aggregationsstufen *immer* auch Sperrungen in den zugehörigen Untertabellen niedrigerer Verdichtung bedeuten. Nur dann, wenn sich ausnahmsweise

Abb. 1.5

Hinaussperren

Kreis (a = Anzahl, b = Betrag)		Gruppe A	Gruppe B	Gruppe C	Gruppe D	Σ
1	a		8			**8**
	b		240			**240**
2	a	3	● ~~2~~	⊙ ~~33~~	67	**105**
	b	240	~~187~~	~~184~~	1 782	**2 393**
3	a		3			**3**
	b		316			**316**
4	a		87			**87**
	b		448			**448**
Regierungsbezirk	**a**	**3**	⊕ ~~**100**~~	⊕ ~~**33**~~	**67**	**203**
	b	**240**	~~**1 191**~~	~~**184**~~	**1 782**	**3 397**

Abb. 1.6

Hineinsperren

Kreis (a = Anzahl, b = Betrag)		Gruppe A	Gruppe B	Gruppe C	Gruppe D	Σ
1	a	11	8	4	117	**140**
	b	7 760	240	57	4 154	**12 211**
2	a	⊕ ~~3~~	● ~~2~~	⊕ ~~33~~	⊙ ~~67~~	**105**
	b	~~240~~	~~187~~	~~184~~	~~1 782~~	**2 393**
3	a	322	⊙ ~~3~~	18	● ~~8~~	**351**
	b	1 723	~~316~~	115	~~258~~	**2 412**
4	a	116	87	21	4	**228**
	b	842	448	439	86	**1 815**
Regierungsbezirk	**a**	□ ~~**452**~~	**100**	□ ~~**76**~~	**196**	**824**
	b	~~**10 565**~~	**1 191**	~~**795**~~	**6 280**	**18 831**

● = geheim zu haltender Wert
⊕ = andere Hierarchie-Stufe erzwingt Sperrung.
□ = Löschung höherer Hierarchie-Stufen
⊙ = Löschung zur Vermeidung der Errechenbarkeit der geheim zu haltenden Werte

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.

1.4.2 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 Aggragationsstufen 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),

Beispieltabelle mit Zwischensummen in zwei Gliederungen

Abb. 1.7

2. Schlüssel

1. Schlüssel	ACD	ACC	ACB	ACA	AC	ABC	ABB	ABA	AB	AAD	AAC	AAB	AAA	AA	A
0000134	112	10	1 445	549	2 116	4 128	345	211	4 684	321	0	0	95	416	7 216
	5 S	2 P	20 S	12 S	39	34	15	12	61	21 S	0	0	2 P	23	123
0000133	40	66	0	23	129	2 567	2 332	432	5 331	732	644	0	0	1 376	6 836
	1 P	4 S	0	3 S	8	44	30	21	95	51	34	0	0	85	188
0000132	723	254	327	543	1 847	1 123	4 427	1 632	7 182	432	0	234	0	666	9 695
	9	11	5	19	44	64	59	26	149	23	0	36	0	59	252
0000131	2 156	1 342	1 111	99	4 708	590	2 334	342	3 266	34	0	0	256	290	8 264
	33	23	17	4	77	11	28	9	48	3 S	0	0	17 S	20	145
0000130	3 031	1 672	2 883	1 214	8 800	8 408	9 438	2 617	20 463	1 519	644	234	351	2 748	32 011
	48 S	40	42 S	38	168	153	132	68	353	98	34	36	19	187	708
0000125	321	11	411	0	743	0	56	0	56	712	3 421	0	0	4 133	4 932
	5 S	3 S	18	0	26	0	5	0	5	50	84	0	0	134	165
0000124	56	12	2 152	399	2 619	0	123	0	123	345	2 612	55	0	3 012	5 754
	4 S	1 P	29	11	45	0	10	0	10	44	61	3	0	108	163
0000123	99	311	754	345	1 509	221	34	73	328	123	321	567	43	1 054	2 891
	8	10	19	16	53	7	2 P	6 S	15	23	41	32	4	100	168
0000122	1 837	19	88	0	1 944	0	621	0	621	1 015	2 221	96	641	3 973	6 538
	33 S	1 P	4	0	38	0	13	0	13	89	52	18	8	167	218
0000121	344	298	0	934	1 576	0	74	0	74	0	231	0	875	1 106	2 756
	15	13	0	9	37	0	8	0	8	0	33	0	72	105	150
0000120	2 657	651	3 405	1 678	8 391	221	908	73	1 202	2 195	8 806	718	1 559	13 278	22 871
	65	28	70	36	199	7	38 S	6 S	51	206	271	53	84	614	864
0000113	53	221	29	1 001	1 304	0	0	0	0	11	0	21	0	32	1 336
	2 P	8	3 S	19	32	0	0	0	0	2 P	0	2 P	0	4	36
0000112	423	0	0	0	423	0	261	34	295	745	0	67	0	812	1 530
	18	0	0	0	18	0	5 S	2 P	7	71	0	8	0	79	104
0000111	28	0	0	0	28	0	0	0	0	148	0	81	0	229	257
	5	0	0	0	5	0	0	0	0	25 S	0	7 S	0	32	37
0000110	504	221	29	1 001	1 755	0	261	34	295	904	0	169	0	1 073	3 123
	25 S	8	3 S	19	55	0	5 S	2 P	7	98	0	17	0	115	177
0000100	6 192	2 544	6 317	3 893	18 946	8 629	10 607	2 724	21 960	4 618	9 450	1 121	1 910	17 099	58 005
	138	76	115	93	422	160	175	76	411	402	305	106	103	916	1 749

Legende: Wert — 10.000
Berichtspflichtige — 100 P Sperrvermerk (P = primär, S = sekundär)

(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 demgemäß 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 Rückrechung 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 Schutzparnter in einem Sicherungsquader ungeeignet. Werden die Fallzahlen nicht veröffentlicht, die zu geheimen Werten beitragen, so kommt der Annahme einer erhöhten Schutzbedürftigkeit die Bedeutung einer Vorinformation über die Tabellenwerte zu, indem 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 Streifen 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: Be-troffen 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 untersster Aggregation. Um bei der Sicherung der Sekundärsperrungen in Zeile 130 möglichst viele bereits gesperrte Tabellenwerte miteinzubeziehen, 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.

1.4.3 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.

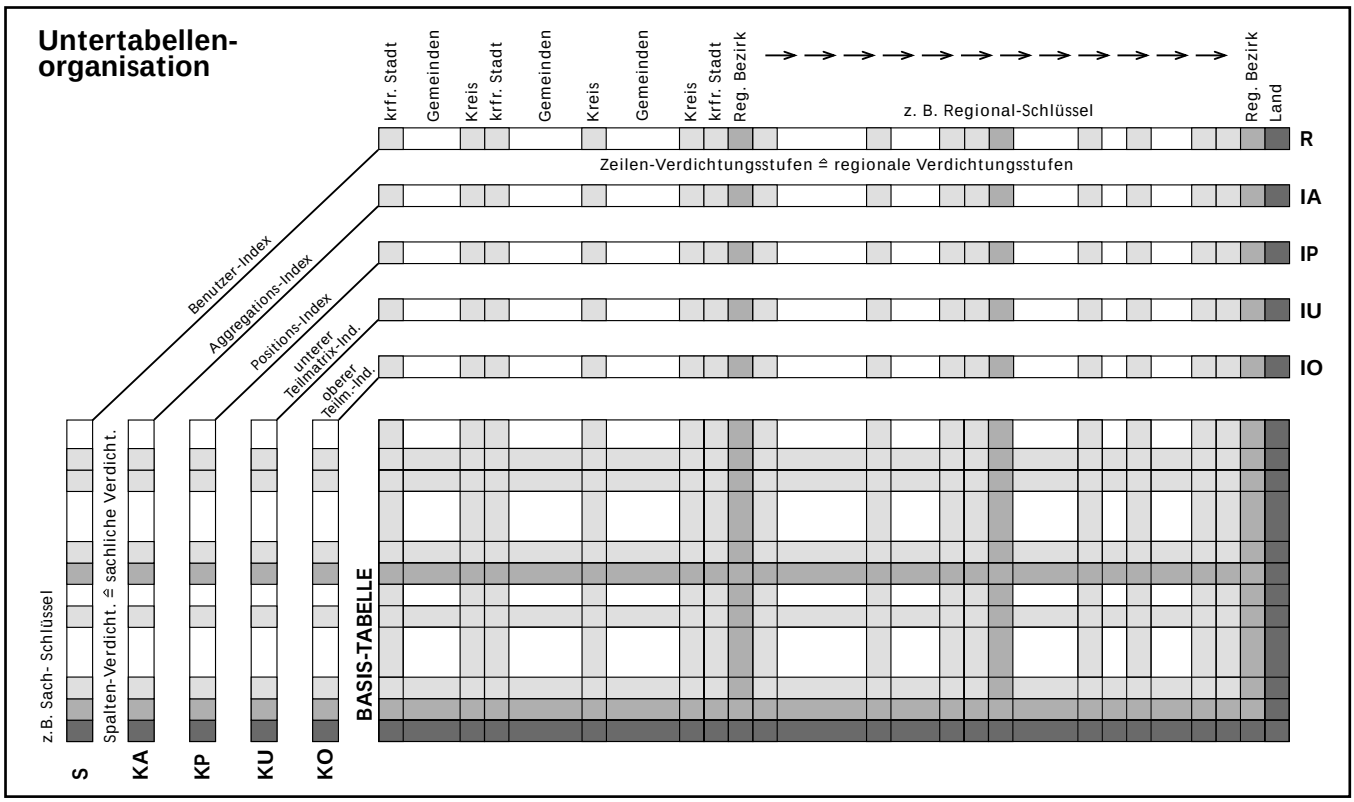
Um dabei alle Untertabellen einer Statistik in Programmschleifen abarbeiten zu können, muss jede einzelne von ihnen als ganzes ansprechbar sein. Als „Ansprechmerkmale“ sind die Aggregationsstufennummern der Gliederungskriterien einer Untertabelle ohne ihre Randsummen (z. B. die Aggregationsstufen der sachlichen und der regionalen Gliederungsmerkmale der Abb. „Eingabe-Daten“ 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 demgemäß 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.

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 Unterta-

Abb. 1.8



bellen 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.

1.5 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 nahe liegende 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 Werte 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 berechnen, indem man von den beiden Zeilengleichungen der zweiten und dritten Zeile die beiden Spaltengleichungen der Spalten B und D subtrahiert.

Abb. 1.9

Kreis a = Anzahl b = Betrag		Gruppe				Σ
		A	B	C	D	
1	a	X_1	8	X_2	117	**140**
	b		240		4 154	**12 211**
2	a	3	X_3	33	X_4	**105**
	b	240		184		**2 393**
3	a	322	X_5	X_6	X_7	**351**
	b	1 723				**2 412**
4	a	X_8	87	X_9	4	**228**
	b		448		86	**1 815**
Regierungsbezirk	**a**	**452**	**100**	**76**	**196**	**824**
	b	**10 565**	**1 191**	**795**	**6 280**	**18 831**

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.

2. Quaderverfahren zur Vermeidung eindeutiger Rückrechenbarkeit geheimer Werte

2.1 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 die primär geheimen Werte einer n-dimensionalen Untertabelle, ob sie einem n-dimensionalen Quader mit lauter gesperrten Werten angehören (Prüffunktion des Quaderverfahrens) und sichert sie 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 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.

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

Zur Einführung des Quadersicherungskonzepts betrachte man zunächst die schematisch dargestellte dreidimensionale Tabelle, in der der durch die Ausprägungen der drei Gliederungskriterien (Indizes (g_1, g_2, g_3)) fixierte geheime Tabellenwert durch weitere geheime Werte (Primär- oder Sekundärsperrungen) in den Ecken eines dreidimensionalen Quaders gesichert worden ist.

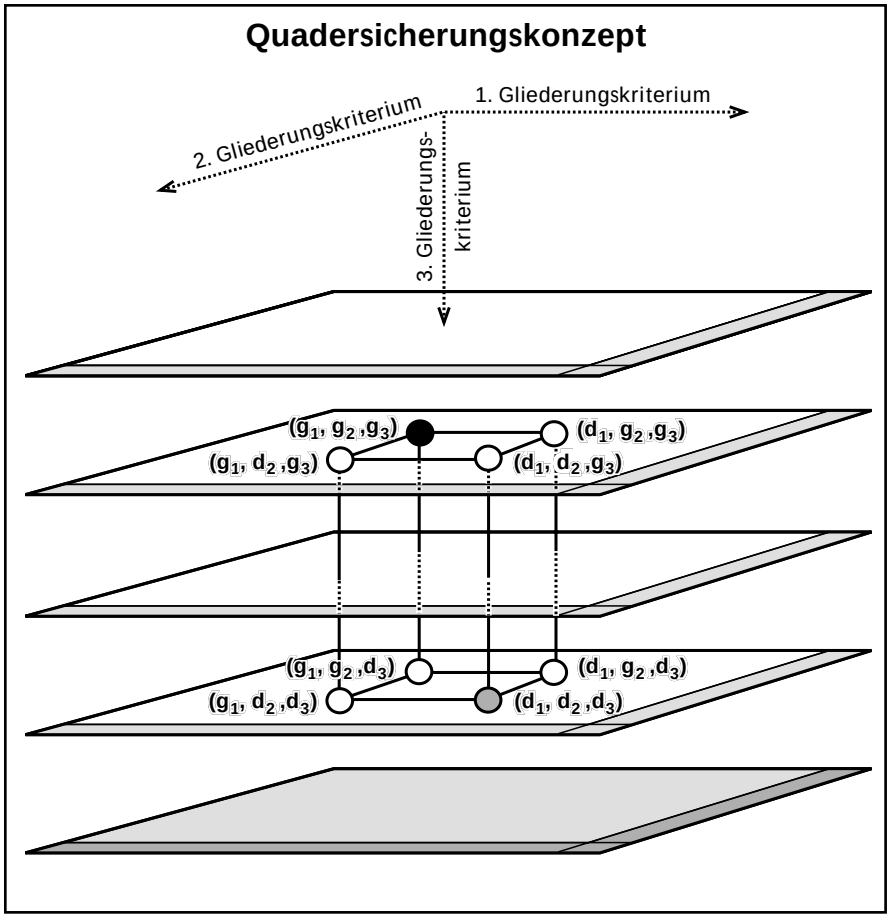
Um den geheimen Wert zuerst in seiner Ebene mit festgehaltenem Gliederungswert g_3 zu schützen, wird das Karree $K(g_3) = \{(g_1, g_2, g_3), (d_1, g_2, g_3), (g_1, d_2, g_3), (d_1, d_2, g_3)\}$ ausgewählt. Da eine Rückrechnung geheimer Werte auch über die dritte Gliederung erfolgen kann, wird noch ein 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

Abb. 2.1



einem anderen diametral, wenn sich die Indizes beider Tabellenwerte in jedem Gliederungskriterium 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 eine möglichst kleine Anzahl von Sekundärsperrungen zu erzielen, werden folgende Regelungen getroffen:

1. **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. **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 jede Einzelangabe des ersten Quaders – mit Ausnahme des zu schützenden Wertes – ausschließt.**

2.1.2 Behandlung von Einzelangaben

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. Dennoch ist Definition 3 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 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 diese Regelung wird vermieden, dass durch eine rigorose (Einzel-)Quadersicherung gemäß Definition 3 in Tabellen der Gestalt der Abb. 2.2 alle Summenwerte gesperrt werden müssen! Die Beispieltabelle ist bereits durch ihre Primärsperrungen gegen Rückrechnung gesichert.

Eine außergewöhnliche Situation bei der Behandlung von Einzelangaben liegt auch 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 Regelungen des vorhergehenden Abschnitts 2.1.1 weder mit der Einzel- noch mit der Doppelquaderauswahl zu befriedigen. 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 wieder-

Abb. 2.2

Kreis / a = Anzahl, b = Betrag		Gruppe A	Gruppe B	Gruppe C	Σ
1	a	●1	●1	●1	3
	b	10	20	40	70
2	a	●1	●1	●1	3
	b	20	30	10	60
3	a	●1	●1	●1	3
	b	30	10	50	90
Regierungsbezirk	**a**	3	3	3	9
	b	60	60	100	220

● = geheim zu haltender Wert

um 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 gleich in mehreren Rändern auftritt, weil sie ja immer als nur ein Tabellenwert in die Betrachtung einbezogen wird.

Der Sonderfall der Einzelangabe im Rand ist nach diesen Überlegungen zwar durch die Regelungen 2.1.1 abgedeckt, die Feststellung aber, ob gewisse Einzelangaben eines n-dimensionalen Quaders zum selben Berichtenden gehören, erfordert zusätzliche Abfragen in einer der „innersten" und damit am häufigsten durchlaufenen Programmschleifen. Um sie zu umgehen, kann man sich folgender sehr einfacher Regel bedienen, die es erlaubt, bereits vor Abarbeitung der betreffenden Untertabelle Mehrfacheingänge von Einzelangaben zum selben Merkmalsträger in *einem* Sicherungsquader zu tolerieren.

Einzelangaben – 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 ist bereits durch den Quader der zugehörigen Einzelangabe im Tabelleninneren geschützt).

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 – anderenfalls wäre der Randsummenwert keine Einzelangabe. Damit verbietet sich also der Einsatz von Randeinzelangaben als Sicherungspartner bei der Einzelquadersicherung, es sei denn zum Eigenschutz desselben nur durch die Indizierung unterschiedenen Wertes.

Es ist hier darauf hinzuweisen, dass obige Regel auch dann ihre Gültigkeit behält, wenn sogenannte Nullwerte als Sicherungspartner in einem n-dimensionalen Quader zugelassen sind (vgl. 3.1.2), wenn diese von einer von Null verschiedenen Anzahl von Berichtenden gemeldet wurden. Lediglich leere Tabellenfelder als Sperrkandidaten könnten Probleme bereiten, weil damit das paarweise Auftreten von Einzelangaben im Rand und im Inneren der Tabelle durchbrochen werden könnte, weil anstelle der Einzelangabe im Tabelleninneren auch noch gewisse leere Felder als Sperrkandidaten zur Auswahl stehen. Um dies zu unterbinden, werden sperrbare leere Tabellenfelder mit sehr großen fiktiven Werten belegt, so dass sie keine Alternative zur Einzelangabe im Rand darstellen (vgl. 5.1).

2.1.3 Abschätzung des Rechenaufwands beim Quaderverfahren

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, Vergleich 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 $N_g \leq m_1 * m_2 * ... * m_n = T$ 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 Kapitel 7).

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, der mit Ausnahme des zu schützenden Wertes keine Einzelangaben enthält, im Tabelleninneren tatsächlich existiert. Die Voraussetzung ist auch schon bei weniger exotischen Tabellen als Abb. 2.2 verletzt, 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 Regelung 2 ein Quaderpaar ausgewählt 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 nicht jede Einzelangabe eine Doppelquaderbildung erzwingt (siehe dazu die Beispieltabelle des Abschnitts 1.4.2); 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, so dass auch bei großen Tabellen wie der Umsatzsteuerstatistik NRW 1994 noch akzeptable Rechenzeiten erreicht werden (siehe vorletzter Abschnitt). 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, sodass der Tabellenumfang in diesen Fällen die Rechenzeit noch verstärkt beeinflusst.

2.2 Herleitung der Quader-Indexformel

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

$$g = (g_1, g_2, \ldots, g_i, \ldots, g_n)$$

gegeben. Ein dazu diametraler Tabellenwert D habe die Indizes

$$d = (d_1, d_2, \ldots, d_i, \ldots, d_n)$$

mit

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

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} mit

$$q = (q_1, q_2, q_3, \ldots, q_i, \ldots, q_n)$$

indiziert sind, wobei gilt (Definition 2):

$$q_i = \begin{cases} \text{entweder} & g_i \\ \text{oder} & d_i \end{cases}, \; i = 1, 2, 3, \ldots, n$$

Da demgemäß 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.

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

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

eine binäre Variable eingeführt wurde und k die Nummer des betrachteten Normquaderwertes bezeichnet – in einer nun herzuleitenden Numerierung –. 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 umcodieren.

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, $k = 01001_{bin} = 0*2^4+1*2^3+0*2^2+0*2^1+1*2^0 = 9_{dez}$. Demgemäß 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 Numerierung 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
$q_i(k) = g_i \leftrightarrow B_i(k) = 0$ ^
$q_i(k) = d_i \leftrightarrow B_i(k) = 1$.

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

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

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-Indexformel setzt voraus, dass die Ausprägungen der Gliederungsmerkmale ganzzahlig sind; anderenfalls kann es hilfreich sein, sich eine temporäre Umindizierung in ganze Zahlen vorzustellen.

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 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 auch Tabellen mit positiven und negativen Werten mit Intervallschutz zu sichern.

Zuerst soll jedoch der Intervallschutz für die etwas einfacher zu handhabenden positiven Tabellen aus dem Quaderkonzept hergeleitet werden.

3.1 Bestimmung der Spannweite geheimer Werte in positiven Tabellen

3.1.1 Ansatz zur Spannweitenberechnung mit Hilfe linearer Optimierung

Mit Hilfe der gegebenen Untertabelle kann der externe Daten-Nutzer ein lineares Gleichungssystem

$$C\,X = B$$

für die r unbekannten geheimen Tabellenwerte

$$X^T = (X_1, X_2, X_3, ..., X_r)$$

aufstellen mit gegebener Koeffizienten-Matrix C und gegebenem Konstanten-Vektor B.

Da dieses wegen sekundärer Geheimhaltung auch unter der Voraussetzung nicht negativer Werte nicht eindeutig lösbar ist, löst er die 2 r linearen Optimierungs-Aufgaben (k = 1, 2, 3, ..., r)

$$\text{Minimiere } X_k$$
$$C\,X = B$$
$$X_i \geq 0 \text{ für } i = 1, 2, 3, ..., r$$

$$\text{Maximiere } X_k$$
$$C\,X = B$$
$$X_i \geq 0 \text{ für } i = 1, 2, 3, ..., r.$$

Auf diese Weise können die möglichen Werte der X_k eingegrenzt werden. Mit Hilfe der Lösungen $\max X_k$ und $\min X_k$ erhält man für jeden geheimen Wert X_k eine Spannweite, $\text{range}_k = \max X_k - \min X_k$ (k = 1, 2, 3, ..., r) .

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

3.1.2 Abschätzung der Spannweite geheimer Werte in positiven Tabellen mit Hilfe des n-dimensionalen Quaders

Mit Hilfe des n-dimensionalen Quaders können auf besonders einfache Weise Spannweiten geheimer Werte bestimmt werden, die höchstens so groß wie die mit linearer Optimierung berechneten sind.

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-Indexformel).

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:

$$\mathbf{X + X' = \Sigma} \qquad \mathbf{(2)}$$

Σ 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 Summtionsindizes 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. Demgemäß 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 Variablensymbole mit langen Indexleisten und gestaltet damit die Formelbilder übersichtlicher und einprägsamer.[1)]

Beispiel:

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 •, 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 derselben 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

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

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

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

sein, damit obige Quadergleichung richtig bleibt.

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: $X+X' = \Sigma_{XX'}$, $X'+Y = \Sigma_{X'Y}$, $Y+Y' = \Sigma_{YY'}$,...,$Z+Z' = \Sigma_{ZZ'}$. 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 Indexumbesetzung 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:

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

so dass schließlich auch

$$\hat{Z} = Z + \varepsilon \quad \text{und} \quad \hat{Z}' = Z' - \varepsilon$$

richtig ist[2)]. Obige Gleichungen gelten also für alle Quaderwerte X, X' mit demselben Schätzfehler ε.

Die gemeinsame Lösung (3) aller durch (2) dargestellten Quadergleichungen enthält genau einen Parameter, den Schätzfehler ε. Bei frei wählbarem ε bedeutet demnach der in Abschnitt 2 definierte Quaderschutz (Pkt. 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 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 Pkt. 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-

1) Die Quaderdefinition ergibt für jedes Gliederungskriterium i mit m_i Ausprägungen unmittelbar Xq_1, ..., q_{i-1}, g_i, q_{i+1}, ..., q_n+Xq_1, ..., q_{i-1}, d_i, q_{i+1}, ..., q_n = Bq_1, ..., q_{i-1}, • , q_{i+1}, ..., q_n–ΣAq_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 Σ über alle j, j = 1, 2, ..., m_{i-1}, $j \neq g_i$, $j \neq d_i$ zu erstrecken ist.

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

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 minX sein können. Sind also negative Schätzwerte auszuschließen, so muss für $\varepsilon = \varepsilon_1 \geq 0$

$$\varepsilon_1 \leq \text{minX'}$$

und für $\varepsilon = \varepsilon_2 < 0$ muss

$$|\varepsilon_2| \leq \text{minX}$$

sein. Das heißt:

$$\hat{X} \in [X - \text{minX}, X + \text{minX'}],$$
$$\hat{X'} \in [X' - \text{minX'}, X' + \text{minX}] \qquad (4)$$

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 Schätzwerte, 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:

$$\textbf{range} = \text{minX'} + \text{minX} \qquad (5)$$

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).

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

∑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''.

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

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 werden 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_1), (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; ∞) 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 mit Einserstufen 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, das einen hinreichenden Intervallschutz bietet:

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

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

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 ist dann kleiner 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.)

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.

Mit der Quaderauswahlformel (6) ist nun sichergestellt, dass jeder primär geheime Wert aus noch offenen Tabellenwerten (der betreffenden Untertabelle) höchstens bis auf seine Spannweite genau berechnet werden kann, wobei diese Spannweite durch den vorgegebenen Prozentwert q bzw. durch einen Absolutwert im Falle primär geheimer Nullen festgelegt werden muss.

Darüber hinaus bietet die Formel die Möglichkeit, auch Quaderelemente mit Wert 0 als Schutzpartner für geheime Werte einzusetzen:

Wird der Sicherungsquader so ausgewählt, dass Null-Werte (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 Anwendung obiger Auswahlformel auch einen hinreichenden Intervallschutz.

Um ganz allgemein einen hinreichenden Intervallschutz zu garantieren – wobei auch Null-Werte 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 die 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 ausfällt. Diese Beschränkung auf den Vergleich der relativen Spannweite des zu schützenden primär geheimen Wertes mit der vorgegebenen Schutz-Schranke 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, könnten im statistischen Mittel nur Quader mit größeren Spannweiten, als für den Schutz des primär geheimen Wertes notwendig, herangezogen werden, weil die zur Sicherung benötigten anderen Werte des Quaders u. U. auch größer als der zu schützende geheime Wert selbst sind.

2. Wenn die sekundäre Geheimhaltung mit Untertabellenabgleich erfolgt, so lässt sich ein hinreichender Intervallschutz zu vorgegebener Mindestspannweite in denjenigen Untertabellen, die Randsummensperrungen erfordern, prinzipiell nicht mehr garantieren: Sperrungen in einer Untertabellenrandsumme werden beim Untertabellenabgleich in einer Untertabelle höherer Hierarchie gesi-

chert. 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 Eckwerte 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 (hochdimensionalen) Tabelle lässt sich mit dem Quaderverfahren ein hinreichender Intervallschutz erreichen.

Wie durch den Untertabellenabgleich eingetragene Vorabinformationen in Form von Schätzintervallen für 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 Sicherung der Beispieltabelle mit Intervallschutz und mit Nullwerten als Sperrpartner

Die in Abbildung 3.0 (siehe Seite 23) eingetragenen Sperrungen wurden unter der Voraussetzung nicht negativer Tabellenwerte bei Intervallschutz mit 125 % relativer Mindestspannweite und unter Einbeziehung von Nullwerten als Sperrkandidaten erzielt. Betrachtet man darin 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[3)] 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

3) 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.

Sicherung der Beispieltabelle mit Intervallschutz und Nullwerten als Sperrpartner

Abb. 3.0

2. Schlüssel

1. Schlüssel	ACD	ACC	ACB	ACA	AC	ABC	ABB	ABA	AB	AAD	AAC	AAB	AAA	AA	A
0000134	112	10	1 445	549	2 116	4 128	345	211	4 684	321	0	0	95	416	7 216
	5 S	2 P	20	12 S	39	34	15	12	61	21 S	0	0	2 P	23	123
0000133	40	66	0	23	129	2 567	2 332	432	5 331	732	644	0	0	1 376	6 836
	1 P	4 S	0	3 S	8	44	30	21	95	51	34	0	0	85	188
0000132	723	254	327	543	1 847	1 123	4 427	1 632	7 182	432	0	234	0	666	9 695
	9	11	5	19	44	64	59	26	149	23	0	36	0	59	252
0000131	2 156	1 342	1 111	99	4 708	590	2 334	342	3 266	34	0	0	256	290	8 264
	33	23	17	4	77	11	28	9	48	3 S	0	0	17 S	20	145
0000130	3 931	1 672	2 883	1 214	8 800	8 408	9 438	2 617	20 463	1 519	644	234	351	2 748	32 011
	48	40	42	38	168	153	132	68	353	98	34	36	19	187	708
0000125	321	11	411	0	743	0	56	0	56	712	3 421	0	0	4 133	4 932
	5 S	3 S	18	0	26	0	5	0	5	50	84	0	0	134	165
0000124	56	12	2 152	399	2 619	0	123	0	123	345	2 612	55	0	3 012	5 754
	4 S	1 P	29	11	45	0	10	0	10	44	61	3	0	108	163
0000123	99	311	754	345	1 509	221	34	73	328	123	321	567	43	1 054	2 891
	8	10	19	16	53	7	2 P	6 S	15	23	41	32	4	100	168
0000122	1 837	19	88	0	1 944	0	621	0	621	1 015	2 221	96	641	3 973	6 538
	33 S	1 P	4	0	38	0	13	0	13	89	52	18	8	167	218
0000121	344	298	0	934	1 576	0	74	0	74	0	231	0	875	1 106	2 756
	15	13	0	9	37	0	8	0	8	0	33	0	72	105	150
0000120	2 657	651	3 405	1 678	8 391	221	908	73	1 202	2 195	8 806	718	1 559	13 278	22 871
	65	28	70	36	199	7	38 S	6 S	51	206	271	53	84	614	864
0000113	53	221	29	1 001	1 304	0	0	0	0	11	0	21	0	32	1 336
	2 P	8 S	3	19	32	0	0	0	0	2 P	0	2 P	0	4	36
0000112	423	0	0	0	423	0	261	34	295	745	0	67	0	812	1 530
	18 S	0 S	0	0	18	0	5 S	2 P	7	71	0	8	0	79	104
0000111	28	0	0	0	28	0	0	0	0	148	0	81	0	229	257
	5	0	0	0	5	0	0	0	0	25 S	0	7 S	0	32	37
0000110	504	221	29	1 001	1 755	0	261	34	295	904	0	169	0	1 073	3 123
	25	8	3	19	55	0	5 S	2 P	7	98	0	17	0	115	177
0000100	6 192	2 544	6 317	3 893	18 946	8 629	10 607	2 724	21 960	4 618	9 450	1 121	1 910	17 099	58 005
	138	76	115	93	422	160	175	76	411	402	305	106	103	916	1 749

Legende: Wert — 10 000
Berichtspflichtige — 100 **P** Sperrvermerk (P = primär, S = sekundär)

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 (siehe Fußnote 3): 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 Sicherungsquaders 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 ganz 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 inidzierten 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 sichernden 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.

3.2 Tabellen mit vorgegebenen Schätzintervallen

3.2.1 Berücksichtigung externer Schätzintervalle der Tabellenwerte bei der Spannweitenberechnung mit dem Quaderverfahren

Bei der Sicherung sensibler Daten in einer Veröffentlichungstabelle muss man bedenken, dass die Tabellen-

nutzer 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 der 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.[4)] 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:

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"

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

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 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:

$$\begin{array}{lll} 50 \le X_1 \le 150 & & 50 \le X_1 \le 150 \\ 40 \le 180 - X_1 \le 120 & \text{oder} & 60 \le X_1 \le 140 \\ 45 \le 190 - X_1 \le 135 & & 55 \le X_1 \le 145 \\ 0{,}5 \le 91 - (190 - X_1) \le 1{,}5 & & 99{,}5 \le X_1 \le 100{,}5 \end{array}$$

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 (wenn er nicht auf die range-Formel (5) des Schutzquaders hätte zurückgreifen wollen)

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

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

und daraus als kleinstes Intervall

$$99 \le X_1 \le 180$$

gefunden mit der relativen Spannweite range = 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.

Das hier nur anhand eines speziellen Beispiels über den Intervallschutz bei Vorliegen von zusätzlicher Information vorgeführte Abschätzverfahren für gesperrte Werte wird nun mit Hilfe des Quaderansatzes für 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. Demgemäß 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 X eingrenzt, so dass für jeden Schätzwert $\hat{X}$ des tatsächlichen Wertes X

gilt:

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

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

4) Kritik am Umgang der statistischen Ämter Kanadas, der USA und Europas mit sensiblen Daten von G. Sande (noch zu veröffentlichen). 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

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). Für alle X, X'∈Q gelten also die Gleichungen (3a) und (3b), wenn die beiden benachbarten Quaderwerte die selbe Aggregationsstufen-Summe 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 demgemäß

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

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. $-(X - X_u) \le \varepsilon \le X_o - X$, $-(X' - X'_u) \le -\varepsilon \le X'_o - X'$, 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 $\varepsilon \le \min (X_o - X)$ mit $X \in Q_g$ und $\varepsilon \le \min (X' - X'_u)$ mit $X' \in Q_u$ ein oberer Schrankenwert $\varepsilon_+ \ge 0$

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

und f*ür die absoluten Beträge der negativen ε-Werte*

$|\varepsilon| \le \min (X'_o - X')$ und $|\varepsilon| \le \min (X - X_u)$
$X' \in Q_u$ $X \in Q_g$

der obere Schrankenwert $\varepsilon_- \ge 0$

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

Wenn eine der beiden Quaderteilgesamtheiten nicht existiert (Quader erstreckt sich über das Eckfeld), 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

$$-\varepsilon_- \le \varepsilon \le \varepsilon_+, \ \varepsilon_-, \varepsilon_+ \ge 0$$

mit der Spannweite

$$\textbf{range} = \varepsilon_+ + \varepsilon_- \qquad \textbf{(8)}$$

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 nicht negativen Werten, wenn man die oberen Intervallgrenzen gegen unendlich gehen lässt 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) demgemäß, $\varepsilon_+ = \min [0,5;40] = 0,5$ und nach (7b) die untere Fehlergrenze ε_- zu $\varepsilon_- = \min [40;0,5] = 0,5$ in Übereinstimmung mit oben aufgeführten direkten Berechnungen. Bei Berücksichtigung von externen Schätzfehlern bestimmt i. A. der kleinste Quaderwert mit seinen im Vergleich zu den anderen Quaderwerten besonders kleinen Abweichungen von den Schätzintervallgrenzen die Quaderspannweite, während bei Berücksichtigung der Positivität der Tabelle allein 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.

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 Schätzfehlerbeträgen verursacht und zwar unabhängig davon, ob das betreffende Schätzintervall symmetrisch ist oder nicht. Die Besonderheiten symmetrischer oder nahezu symmetrischer Schätzintervallgrenzen, für die die obige Tabelle (Abb. 3.1 mit Abb. 3.2) ebenfalls ein Beispiel ist, werden im anschließenden Abschnitt näher erläutert.

3.2.2 Abschätzung der Spannweite geheimer Werte im Falle symmetrischer externer Schätzintervalle mit Einbeziehung von Nullwerten

Um den Intervallschutz mit dem Quaderverfahren auch bei Berücksichtigung von Vorinformationen zugewährleisten, müssten zur Berech-

nung 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 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 seine Abweichung von der unteren oder oberen Schätzintervallgrenze als komplexe Zahl zusammenfasst. Der Tabellenwert wird z. B. dem Realteil, seine Abweichung von einer der 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, $\varepsilon_{min}(X)$, 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:

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

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

was wegen $Q = Q_g \cup Q_u$ zu einem für alle Tabellenwerte des selben 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 mit Hilfe der zuletzt genannten Beziehungen mit Berücksichtigung von Vorinformationen kann jetzt auch optional mit dem EDV-Programm GHQUAR.4 durchgeführt werden, wenn man in das zusätzlich in den Datenbestand eingetragene „Gewichtsfeld" 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 gegebenen Schutzintervalls zum Ausdruck kommt. Die Erfüllung der Beziehung (6) mit 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.

Anders liegen die Dinge bei Verwendung der zuerst angegebenen, exakten Fehlerschrankenformel, weil sich dabei unsymmetrische Quaderintervalle bilden können, $\varepsilon_+ \neq \varepsilon_-$, wovon die eine Schranke verschwinden, die andere aber dennoch von Null verschieden sein kann, die Spannweite also nicht verschwindet, es sei denn Nullen treten sowohl in der gerade indizierten wie auch in der ungerade indizierten Quaderteilgesamtheit auf. *Nullwerte können somit auch bei Vorliegen externer Vorinformation ganz legitime Schutzpartner primär geheimer Werte sein.*

Um bei der Sicherung geheimer Tabellenwerte auch Nullen mit einbeziehen zu können, ohne dabei die externe Vorinformation unberücksichtigt lassen zu müssen, ist die näherungs-

weise Berechnung der Spannweite so zu modifizieren, dass auch wieder unsymmetrische Quaderintervalle möglich werden. Es bietet sich an,

$$\varepsilon_{Q+} = \min[\varepsilon_Q, \min X'] \;,\; X' \in Q_u \quad \text{und}$$
$$\varepsilon_{Q-} = \min[\varepsilon_Q, \min X] \;,\; X \in Q_g \qquad (10)$$

als Schrankenwerte einzuführen, die bei Quadern mit ausschließlich nur positiven Werten keinesfalls größer als die einheitliche Schranke ε_Q sind (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 als „kleinsten" Schätzfehler des Nullwertes in das Gewichtsfeld des Eingabedatenbestandes einzutragen, denn das ist genau diejenige Intervallgrenze der Null, die auch in den exakten Schrankenformeln (7a), (7b) wirksam ist.[5)]

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:

$$\textbf{range} = \varepsilon_{Q+} + \varepsilon_{Q-} \qquad (11)$$

(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 Allgemeinen von Null verschieden, der Quader bietet Intervallschutz – ganz analog zum bisherigen Quaderverfahren bei positiven Tabellen.

3.2.3 Eintragung von Schätzintervallen durch andere Tabellen

Viele wichtige Anwendungsbereiche für ein Quaderverfahren mit Berücksichtigung externer Schätzintervalle erschließt die Frage nach den Quellen solcher Vorinformationen. Dieses Vorwissen wird sich in aller Regel auf zuvor veröffentlichtes Datenmaterial gründen. 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. 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 berechnet werden könnten.

Diese „Vorlauftabellen" werden durch Berücksichtigung von Schätzintervallen bei der Sicherung der aktuellen Tabelle weitgehend mitgesichert, weil in älteren Tabellen gesperrte Werte ungenauere aktuelle Schätzwerte 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 immer dann anzuwenden, wenn der die Daten Veröffentlichende mit anderen bereits veröffentlichten Tabellen für die aktuellen Werte Schätzintervalle berechnen kann!

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 jedenfalls nicht 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

---

5) Bei Vorliegen von Nullwerten gilt dann abweichend von (9) $\varepsilon_Q = \min\,[\min_{X \in Q\backslash 0} \varepsilon_{min}(X),\, X_O\,(0)]$. (**9'**)

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

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

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

mit (9') und (7b) folgt

$$\varepsilon_Q = \min\,[\min_{X \in Q_g} \varepsilon_{min}(X),\, \min_{X' \in Q_u\backslash 0} \varepsilon_{min}(X'),\, X_O\,(0)] \le \min\,[\min_{X \in Q_g}(X - X_u),\, \min_{X' \in Q_u\backslash 0}(X_O' - X'),\, X_O\,(0) - 0] = \varepsilon_-,$$

so dass $\varepsilon_{Q-} \le \varepsilon_-$ ist.

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

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 Eintragungen „p“ bzw. „s“ kennzeichnen die primär- bzw. sekundär geheimen Werte.

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)

Abb. 3.8

X_1	0	$99 \le X_1 \le 180$
X_2	89	$10 \le X_3 \le 91$
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:

$0 \le X_1 \quad \wedge \; 99 \le X_1 \le 180$

Zweite Zeile:

$0 \le 101 - X_1 \; \wedge \; 10 \le 190 - X_1 \le 91$

$\Rightarrow \; 99 \le X_1 \le 101$

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).

Um die Sicherheit von geheimzuhaltenden Werten, die nur mit Quadern im Überlappungsbereich zu schützen sind, wesentlich zu verbessern, sollte ganz allgemeinen bei jedem Tabellenabgleich überlappender Tabellen eine Übertragung von Schutzintervallen als externe Schätzintervalle bei der laufenden Tabellensicherung erfolgen. Das kann analog zur Behandlung von Schätzintervallen als Vorinformation geschehen, durch Eintragung der relevanten Abstände der Werte von den jeweiligen Schätzintervallgrenzen in die Imaginärteile der komplexen Tabellenwerte. Ist für einen Wert bereits ein Abstandswert eingetragen, so ist dieser Wert durch den neuen entsprechenden Wert nur dann zu ersetzen, wenn der neue kleiner als der zuvor eingetragene ist. – 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 irrele-

vant, 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). 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 nur einmal durch ein noch stärker eingrenzendes 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 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: $0 \le X_1 \quad \wedge\ 99 \le X_1 \le 180$

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

$$\Downarrow$$

$$99 \le X_1 \le 180$$

$$I_1 = [99; 180]$$

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:

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

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

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

Mit der Quaderschätzwertformel für die gerade indizierten Quaderwerte erhält man mit (3) $\hat{X} = X + \varepsilon$ 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 Pivotelement 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.

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

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 darunterliegend) 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 minX = 39, der kleinste ungerade indizierte minX' = 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 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 „Übersicht zur Struktur des Wertefeldes ...)!

Diese Verschlüsselung lässt sich, wie in der neuesten GHQUAR-Version für den (Unter-)Tabellenabgleich vorgesehen, 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 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

Abb. 3.11 Übersicht zur Struktur des Wertfeldes der Gesamttabelle im Hauptspeicher

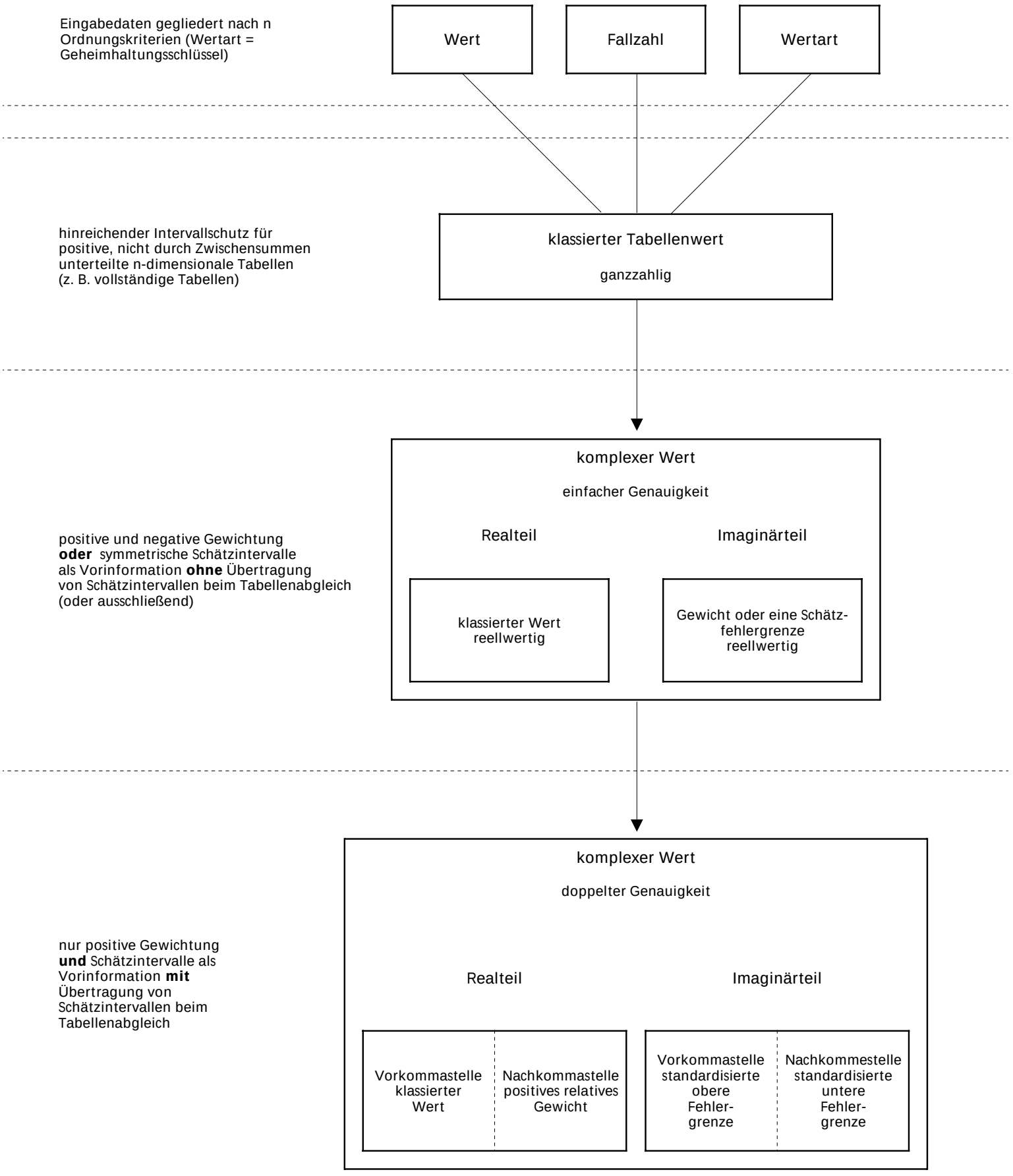


Tabelle als vollständige Tabelle angestrebt werden (vgl. Abschnitt 6).

Bei den bisher realisierten EDV-Programmen werden Schutzintervalle und Spannweiten im Überlappungsfall noch ohne Übertragung von Schätzintervallen berechnet. Von einer möglichen Überschätzung des Intervallschutzes sind aber 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.

Anmerkungen:

1. Anders als bei der Behandlung von Tabellen ohne externe Vorinformationen tritt durch die Eingrenzung der Tabellenwerte 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 nur noch im Bereich von beispielsweise 100 % Abweichung möglich sind.

2. Von praktischer Bedeutung ist auch, dass mit der Einführung von Vorinformation nicht mehr alle offenen Werte als Sperrkandidaten in Betracht kommen, weil ihre Schätzintervallschranken sie zu genau eingrenzen. Durch Einführung von Vorinformation kann somit verhindert werden, dass der Tabellennutzer auf Grund der u. U. engen Eingrenzung von Tabellenwerten durch die Vorinformation primär geheime Werte zu genau berechnen kann.

3. Anders als bei der Behandlung von Tabellen mit sowohl positiven als auch negativen Werten ohne Berücksichtigung von Vorinformation 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 die Auswahl von Sicherungsquadern eingegrenzt, ganz so, wie bei positiven Tabellen.

Erweiterungen und Anwendungen

4. Anmerkungen zur 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 (vgl. 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 Fehler-

größe 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! Demgemäß beziehen sich Umbuchungsvorgänge – anders als die Quaderwerteverfälschungen – immer nur auf im Quader benachbarte Werte gleicher Aggregation. Hier werden vereinfachend Quader betrachtet, die ganz im Inneren einer Untertabelle liegen.

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

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

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

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

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 Quadereckwerte 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 wäre 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 aber 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 und Sensitivitätsmaße

Die folgende Diskussion der Dominanz und der Sensitivität bezieht sich ausschließlich auf so genannte positive Tabellen, d. h. auf Tabellen, die keine negativen Tabellenwerte enthalten.

In der amerikanischen Literatur, z. B. L. H. COX 1981, findet man eine Definition der so genannten Sensitivität S(X) eines Tabellenwertes X, die ein Maß für die Schutzbedürftigkeit von X ausdrückt: S(X) ist als gewichtete Summe der k größten zu X beitragenden Werte abzüglich des gesamten Summenwertes X darstellbar:

$$S(X) = \sum_{i=1}^{k} W_i * X_i - X, \quad W_i > 0 \qquad (12)$$

Ein Tabellenwert X ist danach sensitiv (schutzbedürftig), wenn S(X)>0 ist. Im Falle k = 1 und W_1 = 100/p, wobei p den Prozentwert der Dominanzschranke der primären Geheimhaltung bezeichnet, stimmt diese Regel mit der Dominanzregel in der primären Geheimhaltung überein, wenn man unter „Dominanz" das Überwiegen des Beitrages nur eines Berichtenden versteht. Der allgemeinen Sensitivitätsdefinition (12) liegt die Vorstellung zugrunde, dass bis zu k – 1 dominante Merkmalsträger in Absprache miteinander ihren Wert bestimmen könnten.

Für das Quadermodell besteht grundsätzlich kein Problem, auch die Sensitivität S (X) als Quaderauswahlkriterium einzuarbeiten, indem einfach nur solche Quader zur Sicherung sensitiver Werte zugelassen werden, für deren benachbarte, d. h. zur selben Randsumme beitragende Wertepaare immer

$$S(X + X') \le S(X) + S(X') \le 0$$

gilt. Alle Summen X + X' des betreffenden Quaders sind dann nicht sensitiv, d. h. es gibt keinen dominierenden Quaderwert. Eine besonders Rechenzeit sparende Variante dieses Modells würde die Sensitivität S (X) neben dem Tabellenwert X selbst in den Tabellenbestand eintragen und bei der Quaderauswahl gemäß o. g. Bedingung berücksichtigen.

Obige „Dreiecksungleichung" liefert allerdings eine für die Quadersicherung recht grobe Abschätzung: Danach ist stets die Summe der Sensitivitäten zweier primär geheimer Werte positiv, die Summe der Werte selbst erscheint sensitiv, d. h. nach dem Sensitivitätskriterium ungeschützt. Für die Sicherung eines primär geheimen Wertes kämen daher als benachbarte Werte immer nur nicht sensitive Werte (mit negativen Sensitivitäten) in Frage; sensitive primär geheime Werte könnten sich gemäß obiger Abschätzung niemals gegenseitig schützen. Quader mit zwei benachbarten Primärsperrungen wären als Sicherungsquader unbrauchbar.

Die Situation verbessern kann eine genauere Abschätzung der Sensitivität, die die vorhandene Information der Tabelle vollständig ausnutzt, also auch die Tabellenwerte mit einbezieht. Beispielsweise ergibt sich bei Berücksichtigung nur eines dominierenden Wertes, d. h. im Falle k = 1, für die Sensitivität der Summe zweier benachbarter Quaderwerte X,X' die exakte Formel S (X + X')= S(X) – X', wenn S(X) – X' größer als S(X') – X ist, anderenfalls hat man in dieser Sensitivitätsformel nur X mit X' zu vertauschen.[6)]

Steht also für jedes Tabellenfeld sowohl der Wert als auch dessen Sensitivität zur Verfügung, so hat man im Fall der Einzeldominanz alle Merkmale im Zugriff, um daraus die Sensitivität der Summe zweier benachbarter Quaderwerte exakt zu berechnen. Zwei in einem Quader benachbarte Werte sind in ihrer Summe demnach nicht sensitiv, wenn jede der (auch positiven) Sensitivitäten der beiden Einzelwerte kleiner als der jeweils benachbarte Wert ist, was auch für zwei sensitive benachbarte Werte zutreffen kann.

4.2.1 Sensitivität und Einzeldominanz

Demgegenüber lässt sich die Sicherung geheimer Tabellenwerte bei Berücksichtigung nur eines dominierenden Wertes mit dem sehr viel anschaulicheren Intervallschutzansatz auch ohne zusätzliche Sensitivitätsangaben unmittelbar mit obigem Sensitivitätsmaß bzw. mit unserer zunächst nur im Falle der primären Geheimhaltung daraus abzuleitenden Einzel-Dominanzregel in Zusammenhang bringen. Dazu braucht man nur die Spannweite des zur Sicherung des primär geheimen Wertes X auszuwählenden Quaders gemäß (6) durch

$$\textbf{range} > X / p * 100, \; p \text{ in } \%, \quad 0 < p < 100 \qquad (13)$$

festzulegen[7)], wobei p den für die primäre Geheimhaltung angesetzten prozentualen Dominanzanteil bezeichnet. (13) gilt nur für zu schützende geheime Werte X, die echt kleiner sind als die Quadersumme Σ, zu der sie beitragen, anderenfalls dominiert der zu schützende geheime Wert die Quadersumme immer. Bei Dominanzschutz dürfen Nullen hinsichtlich ihrer Indizierung immer nur zur selben Quaderteilgesamtheit gehören wie der zu sichernde Wert selbst! Diese Einschränkung in der Quaderauswahl hat ihre Ursache in einer durch (13) festgeschriebenen relativen Mindestspannweite von über 100 % (vergleiche 3.1.3 Sicherung der Beispieltabelle mit Intervallschutz und Nullen als Sperrpartner ... und relativer Mindestspannweite von 1.25).

Zur Begründung betrachte man ein Schutzintervall mit X_u als unterer und X_o als oberer Intervallgrenze eines zu sichernden Wertes $X \in [X_u, X_o]$, für dessen Spannweite obige Ungleichung gilt und für die – wegen der Nichtnegativität der Tabellenwerte – außerdem noch folgende Abschätzung möglich ist:

$$\text{range} = X_o - X_u \le X_o \le \Sigma.$$

Diese Abschätzung liefert mit obiger Bedingung (13) für die Quaderspannweite die "Dominanzregel" für den Anteil, den der zu sichernde Tabellenwert X an der Quadersumme Σ nicht überschreiten darf

$$X/\Sigma < p/100.$$

6) Aus S(X) – X'>S(X') – X bzw. (100/p ∗ X_1 – X) – X'> (100/p ∗ X'_1-X') – X folgt X_1>X'_1 und damit S(X + X') = 100/p ∗ X_1 – X – X' = S(X) – X'.

7) Die angegebene range-Abschätzung läßt sich aus (6) gewinnen, indem man dort q/100 durch 100/p ersetzt; seine Rechtfertigung erfährt dieser Ansatz aber erst durch die nachfolgenden Abschätzungen.

Demnach weist auch S(Σ) jede Quadersumme Σ, zu der X als größerer der beiden Quaderwerte beiträgt, gemäß der üblichen Definition von S (siehe z. B. D.A. Robertson, Luxemburg 1994)

$S(\Sigma) := 100/p * X - \Sigma < 0$ (14)

als nicht sensitiv aus, das heißt, X dominiert keine der Quadersummen, ist also sowohl bezüglich eines Sensitivitätsmaßes als auch bezüglich eines Dominanzkriteriums akzeptabel geschützt.

Dabei hat man aber zwischen zwei Sensititvitätstypen von benachbarten Quaderwerten X und X' zu unterscheiden: $S(X + X')$ bezeichnet die eingangs mit (12) eingeführte Sensitivität der einzelnen Berichtsfällen zuzuordnenden Werte in X und X', während $S(\Sigma)$ gemäß obiger Definitionsgleichung (14) die vergröberte, direkt mit der Quaderspannweite zusammenhängende Sensitivität der aggregierten Tabellenwerte X und X' bedeutet, die zur Quadersumme Σ beitragen.

Bei Berücksichtigung nur eines dominierenden Wertes X_1 bzw. X'_1 von zu schützenden Einzelangaben in einem Aggregat X bzw. X', d. h. im Fall von Einzeldominanz ist immer $S(\Sigma) \geq S(X + X')$, weil für $X' < X$ (o. B. d. A.) $\max(X_1, X'_1) \leq X$ und somit $S(X + X') = 100/p * \max(X_1, X'_1) - (X + X') \leq 100/p * X - \Sigma = S(\Sigma)$(*). D. h. bei Einzeldominanz garantiert ein Sicherungsquader für den zu schützenden Wert X mit Quaderspannweite range > X/p * 100 in allen Nachbarschaftspaarungen von X, in denen X der größere Wert ist, immer auch den Schutz des einzelnen dominierenden Wertes X_1 innerhalb des zu sichernden Tabellenwertes X.

Wenn man (13) nicht nur für den zu schützenden Wert, sondern für alle Quaderwerte X des betreffenden Sicherungsquaders fordert, sind wegen $100/p * X - \Sigma < 0$ und obiger Abschätzung (*) alle Quaderwerte hinsichtlich ihrer Einzeldominanz bzw. auch hinsichtlich ihrer Sensitivitäten geschützt. In den bisherigen Anwendungen wurde aber nur die Einhaltung von (13) für den einen zu sichernden Quaderwert gefordert; dann kann im Fall von Einzeldominanz allerdings auch nur erreicht werden, dass der zu schützende Wert wie auch der dazu beitragende größte Einzelwert in keiner Summe dominiert, während der jeweilige Partnerwert und sein größter Einzelwert durchaus dominant sein können: Diese Aussage ist auch für den bisher nicht betrachteten Fall richtig, bei dem der Nachbarwert X' größer als der zu schützende Wert X ist; falls dann der größte Einzelwert X_1 von X größer als der größte Einzelwert X'_1 von X' ist, folgt aus der Abschätzung $100/p * X_1 - \Sigma \leq 100/p * X - \Sigma < 0$ (gemäß (13)), dass weder X noch X_1 dominiert, im Fall $X_1 \leq X'_1$ ist diese Aussage für X_1 trivial.

4.2.2 Sensitivität und Zweifachdominanz

Bei mehr als einem bei der Geheimhaltung zu berücksichtigenden dominierenden Einzelwert genügt die Bedingung (13) nicht, um außer für einen hochgradigen Intervallschutz zu sorgen, auch die Sensitivität bzw. die Mehrfachdominanz mit hinreichender Sicherheit zu behandeln. Als Gegenbeispiel betrachte man eine Einzelangabe X, die einem primär geheimen Tabellenwert X' mit einem dominierenden Wert $X'_1 > 0$ in einem Quader benachbart ist. Wenn dann $X > X'$ ist, hat man für die Sensitivität gemäß (12) $S(X + X') = 100/p * (X + X'_1) - X - X'$ anzusetzen, sodass immer $S(X + X') > S(\Sigma)$ ist. Zum Schutze von mehr als einem Dominierenden reicht also die Bedingung $S(\Sigma) \leq 0$ nicht mehr aus, die Spannweitenbedingung (13) allein garantiert keinen Schutz gegen Mehrfachdominanz.

Beim Schutz von dominierenden Angaben kommt neben der Einfachnoch der Zweifachdominanz, wo zwei Werte zu ihrem Tabellenwert den überwiegenden Beitrag leisten, eine ganz besondere Bedeutung zu: Bliebe die Zweifachdominanz unberücksichtigt, könnte jeder der beiden Dominierenden den gemeldeten Wert des jeweils anderen mit einer Genauigkeit abschätzen, die nur durch die Summe der restlichen kleineren Werte in diesem Tabellenfeld beschränkt ist. Mehr als zwei Dominierende könnten nur durch Absprache miteinander einen der dominierenden Werte in einem Tabellenfeld abschätzen; dieses Problem ließe sich gemäß einer privaten Mitteilung von Frau Gießing, Statistisches Bundesamt, ggf. durch Zusammenfassung der für solche „Kartelle“ in Frage Kommenden auf den Fall der Zweifachdominanz zurückführen. Im folgenden wird daher der Schutz der Zweifachdominanz mit dem Quaderverfahren noch eingehender behandelt.

Ausgangspunkt für die Diskussion der Zweierdominanz ist wieder die hier besonders hilfreiche Definition der mit der Quaderspannweite gemäß (13) zusammenhängenden vergröberten Sensitivität (14), weil durch diese Sensitivitätsformel eine EDV-mäßig leicht zu bearbeitende Fallunterscheidung iniziiert wird: Die Bearbeitung der Zweifachdominanz erfolgt mit der Sensitivitätsformel (14), falls die in der Summe benachbarter Werte, X + X' , dominierenden beiden Einzelwerte X_1 , X'_1 in ihrer Summe kleiner oder höchstens gleich einer der beiden benachbarten Werte X , X' ist, oder sonst mit (12). D. h., nur wenn die Summe aus dem größten Einzelwert X_1 , der zu X beiträgt, und dem größten Einzelwert X'_1 , der zu X' beiträgt, größer als jeder der beiden benachbarten Werte X , X' ausfällt, wird $S(X + X')$ mit (12) exakt berechnet, sonst mit (14) abgeschätzt.

Fordert man, dass (13) für alle Werte eines für die Sicherung geheimer Werte geeigneten Quaders gilt, so kommt es demnach nur darauf an, ob die Summe der beiden größten Einzelwerte, $X_1 + X'_1$, je zweier benachbarter Quaderwerte X , X' größer ist als jeder Einzelne der Nachbarwerte oder nicht. Dazu muss für jedes Tabellenfeld, außer dem Wert selbst, zusätzlich nur noch ein weiterer Wert, der jeweils größte Einzelwert, bereitgestellt werden – das kann in der doppeltgenauen komplexen Schreibweise durch Doppelbelegung des dem Gewicht zugeordneten Anteils im Realteil geschehen (vgl. S. 28f.), indem die drei ersten Stellen hinter dem Komma dem Gewicht, die dann folgenden Stellen X_1 zugeordnet werden –. Ist dann die

Summe der größten Einzelwerte zweier benachbarter Tabellenwerte größer als jeder der Nachbarwerte, so berechnet man die Sensitivität nach der Formel $S(X + X') = 100/p * (X_1 + X'_1) – X – X'$ exakt und beurteilt sie auch danach, anderenfalls gilt unter Verwendung der vergröberten Sensitivität $S(\Sigma)$ die Abschätzung $S(X + X') \leq S(\Sigma)$. Hinsichtlich des Sensitivitätsschutzes bei Berücksichtigung zweier dominierender Einzelwerte wird demnach ein Quader als Sicherungsquader akzeptiert, wenn im ersten Fall für jedes Paar benachbarter Werte $S(X + X') \leq 0$ ist oder wenn im zweiten Fall $S(\Sigma) \leq 0$ ausfällt, sonst wird er verworfen.

Die anschauliche Überleitung von der Quaderspannweite ‚range' zur Sensitivität darf nicht darüber hinwegtäuschen, dass beide Sicherungskriterien grundverschieden sind. Die Verschiedenheit beider Kriterien wird durch das Nullen-Problem besonders deutlich: Während das range-Kriterium Quader mit Nullen als Sicherungsquader akzeptiert, wenn die Nullen in nur einer der beiden Quaderteilgesamtheiten auftreten, ist ein Sensitivitäts- oder Dominanzschutz mit Nullen höchstens dann zu machen, wenn diese Nullen dem zu sichernden Wert nicht benachbart sind. Beim Hinzufügen einer Null zu einem Wert mit dominierenden Einzelwerten bleibt die Sensitivität gemäß (12) positiv, der dominierte Wert wird durch die Addition der Null nicht geschützt. Darüber hinaus erfordert die Vermeidung der „vergröberten" Dominanz von Tabellenwerten gemäß (13) die Berücksichtigung folgender

Sensitivitätsregel:
Quader zur Sicherung der Sensitivität bzw. zur Unterbindung der Dominanz von Einzelwerten in einem Tabellenwert dürfen Nullen höchstens dann enthalten, wenn diese genauso indiziert sind, wie der geheime zu sichernde Wert selbst!

Die Nullen werden aus der Gesamtheit der Sekundärsperrungen noch stärker verdrängt, wenn – wie eingangs dargestellt – alle Paare in einem Sicherungsquader benachbarten Werte gegen Dominanz geschützt werden sollen. Doch auch dann gibt es Beispiele für akzeptable Sicherungsquader mit Nullen.

Diese die allgemeinen Ergebnisse des Abschnitts 3 über die Verteilung von Nullen in Sicherungsquadern stark einschränkende Aussage liegt in dem unterschiedlichen Vorwissen begründet, das man bei der Konstruktion beider Kriterien unterstellt hat: Verfügt der Tabellennutzer über ein Vorwissen in Gestalt von Schätzintervallen, die die Tabellenwerte überdekken, oder ist ihm wenigstens bekannt, dass es sich bei den Tabellenwerten um nicht negative Werte handelt, so ist ein Intervallschutzverfahren mit einer relativen Mindestspannweite kleiner als 1 angezeigt, das verhindert, dass er die geheimen Werte mit den anderen noch offenen Tabellenwerten zu genau berechnen kann; stehen dem Tabellennutzer aber Schätzwerte über die Anteile zur Verfügung, mit denen gewisse besonders große Einzelwerte zu ihren Tabellenwerten beitragen, so muss man durch Addition von Null verschiedener (im Quader dem zu schützenden Wert benachbarter) Werte das bekannte oder näherungsweise bekannte Anteilsverhältnis in ein in der Summe beider Werte Unbekanntes überführen. Zum Vorwissen dieser Art gehört auch die Kenntnis, ob es sich um eine Einzelangabe in einem Tabellenfeld handelt oder ob dort nur zwei Einzelangaben zusammengefasst wurden – beides wären primäre Geheimhaltungsfälle, die auch nicht durch Nullen zu schützen wären, es sei denn, es sollte „nur" Intervallschutz gewährleistet sein.

Intervallschutz betrifft also weitgehend die Sicherung eines geheimen Wertes als Ganzes bezüglich seiner Tabellenumgebung. Dominanzschutz ist auf die Information über das „Innere" eines Tabellenwertes gerichtet.

In aller Regel wird der externe Tabellennutzer über beide Arten von Vorinformationen verfügen, über Schätzintervalle für die Tabellenwerte, weil er vor dem Erscheinen der Tabelle vergleichbare Tabellen ausgewertet hat, und er kennt gewisse Anteile von großen Werten an den Tabellenwerten, weil er z. B. selbst zum Kreis der Berichtspflichtigen gehört. Es wird daher auch im Falle der Mehrfachdominanz, wie bisher bei der Einfachdominanz praktiziert, eine Kombination von range- und Sensitivitätskriterium anzuwenden sein.

Um mit dem Quaderverfahren den geforderten Dominanzschutz für den zu sichernden Wert X durch seine im Quader benachbarten Tabellenwerte X' bis zur Zweifachdominanz zu gewährleisten, genügt es, die Bedingung (13) für nur diesen Wert X bzw. für X' , falls X' > X ist, zu erfüllen und für alle in dem betreffenden Quader dem Wert X benachbarten Werte X', deren Summe größter Einzelwerte $X_1 + X'_1$ größer als jeder der beiden Quaderwerte ist, die Sensitivität $S(X + X') = 100/p * (X_1 + X'_1) – X – X'$ zu berechnen. Ist (13) nicht erfüllt oder eine dieser Sensitivitäten positiv, wird der Quader als Sicherungsquader verworfen, sonst wird er akzeptiert. In $S(X + X')$ bzw. in (13) kann auch ein anderer p-Wert als bei Einzeldominanzprüfung eingetragen werden.

5. Justierung der Verteilung von Sekundärsperrungen nach externen Vorgaben

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 Distributors 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. Dabei wird die zu minimierende Summe zu sperrender Werte des jeweils zur Auswahl stehenden Quaders 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 des EDV-Programms GHQUAR unvermeidbar ist, weil dabei auch die Speicherplatzorganisation eine wichtige Rolle spielt.

5.1 Justierung der Auswahl von Sicherungsquadern durch 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 die Nachweisungsfälle und die von diesen berichteten Werte ausweisen, ist die zu sperrende Wertesumme 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 Beipiel – 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 der letzten Programmversion GHQUAR 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 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. Die 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 fehlerhafter 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 im EDV-Programm GHQUAR 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.

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 Ganze-Zahlen-Achse entsteht eine disjunkte hierarchische Staffelung von Klassenwerten. Alle Klassenwerte einer betrachteten Hierarchiestufe (Staffel) sind stets größer als alle Klassenwerte der niedrigeren Hierarchiestufen. Mit dieser Hierarchiestaffellung 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 Tabellen-(Klassen-)Werte 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 im EDV-Programm GHQUAR 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 Schrankenwerte sind 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 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. Bezeichnet K_{max} den größten, $K_{min} \geq 0$ den kleinsten Klassenwert der noch offenen Tabellenwerte und Kg den einem geheimen Tabellenwert in der Quadersumme zuzuordnenden Wert, so ist obige Forderung erfüllt, wenn die Ungleichung

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

gilt und dies trifft jedenfalls zu, wenn

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

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 Kg_1 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 (wie es die Regelungen vom Abschnitt 2 eigentlich vorschreiben). 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, wo der ursprüngliche klassierte Tabellenwert in den Realteil, sein Gewichtswert in den Imaginärteil der komplexen Zahl eingetragen wurde: Die inzwischen so erweiterte EDV-Programm-Version GHQUAR, Version 4, bietet nun die Möglichkeit einer freien kontinuierlichen Gewichtung, ohne dabei auf den originalwertebezogenen Intervallschutz verzichten zu müssen.[8)] Die Liste der verfügbaren Justierungsmaßnahmen ist nun noch um einen dritten Punkt zu ergänzen.

5.3 Justierung durch externe Gewichtung

GHQUAR, Version 4 sieht im Satzformat des Eingabebestandes ein zusätzliches nummerisches Tabellenfeld vor, in das ein beliebiger reeller Zahlenwert (zwischen -10^{50} und $+10^{50}$) eingetragen werden kann, mit dem dann bei der Berechnung der Quadersumme die offenen klassierten Werte multipliziert werden. Ein gewichteter offener Wert ist als Sicherungspartner in einem zweidimensionalen Quader attraktiver als ein primär geheimer Wert (einschließlich Einzelangaben), wenn sein „Gewicht" kleiner als –330 000 ist (bei höherdimensionalen Tabellen erhöht sich der Betrag dieses Schrankenwertes entsprechend der Zunahme der Quaderwerte, siehe 5.2.2), während ein sehr großer Gewichtswert, z. B. 10^{50}, einen offenen Tabellenwert vor einer Sekundärsperrung weitgehend bewahrt. Eine exakte Aussage, ab welcher Gewichtsgröße ein offener Tabellenwert mit Sicherheit offen bleibt, ist im Allgemeinen 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 überschaubar 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) Gleichwertigkeit bei der Auswahl von zu sperrenden Tabellenwerten (nach dem Summenkriterium) wird bei posititven Tabellen erreicht, wenn man für die Tabellenwerte deren reziproken Logarithmus aus dem auf den Tabellenminimalwert bezogenen Wert als Gewichtsfunktion verwendet und diese Werte als Gewichte in den Eingabebestand von GHQUAR einträgt. 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, wie groß der ist.

(b) Soll die Anzahl der Berichtenden die Sperrpositionen mitbestimmen, wobei auch der Betrag des Tabellenwertes von Einfluß ist, so bietet sich als Gewichtsfunktion die Anzahl der Berichtenden an; soll aber die Anzahl der Berichtenden *allein* das Ausmaß des Informationsverlustes durch die Sekundärsperrungen 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 dabei auch 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 Pkt. 5.2.1 – unterschiedlich für verschiedene Gliederungskriterien oder auch für verschiedene Gliederungsmerkmalsgruppen vorgegeben werden kann. Ein anderes Beispiel für die tabellenfeldpositionsab-

8) In neueren Entwicklungen von GHQUAR werden doppelt genaue komplexe Zahlen als Tabellenwertvariable benutzt, die im Realteil den klassierten Wert, die Gewichtung und den größten Einzelwert aufnehmen; der Imaginärteil speichert die obere und die untere Schätzfehlergrenze (siehe auch Abschnitt 3.2.3). Alle folgenden Betrachtungen beziehen sich auf eine GHQUAR-Version mit komplexer Wertevariabler einfacher Genauigkeit.

hängige 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 auseinanderklaffen, weil sonst betragsmäßig besonders kleine Summanden mitunter gar nichts mehr zur Unterscheidung der Quadersummen beitragen. Das Summenkriterium ist dann in Bezug auf die betragsmäßig kleinen 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 restliche gewichtete 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 rein 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, etwa im Bereich von 1 bis 100, weil dabei Rundungsverfahren noch keinen wesentlichen Einfluss haben.

5.3.2 Externe Gewichtung zur Bearbeitung von Zeitreihentabellen mit dem Quaderverfahren

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.

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. Gängige Verfahren sind bisher die separate querschnittsmäßige Behandlung mit einschlägigen Geheimhaltungsverfahren, d. h. ohne Berücksichtigung des zeitlichen Zusammenhanges, 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 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 Sperrpositionenauswahl 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 geheimgehaltenen 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 Ab-

hängigkeit 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:

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

Dabei bezeichnen

- n = Tabellendimension
- $q_1, q_2, \ldots, q_n$ = Koordinaten des Quaderwertes Q
- $g_1, g_2, \ldots, g_n$ = Koordinaten des Pivot-Wertes G
- $p_1, p_2, \ldots, p_n$ = 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:

1. Sollen alle Gliederungskriterien hinsichtlich ihres Abstandsmaßes gleichstark in die Gewichtung eingehen, so ist
   $$p_1 = p_2 = \ldots = p_n = p$$
   zu setzen; insbesondere ergibt p = 2 das Quadrat des Euklidischen Abstandsmaßes.

2. Positive Werte der p_j erhöhen das Gewicht (verringern also die Sperrwahrscheinlichkeit von Werten mit größeren Koordinatendifferenzbeträgen), negative pj 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 wie z. B. GHQUAR 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 – „überlappende“ Tabellen

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 Einzeltabel-

len 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 hier jedoch 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. 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 Abschnitts 7.

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. 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; 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 jeweiligen 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 der 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 gröbere, das zweite das feinere 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 Geheimhaltungsprüfung 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 obigen Bemerkungen dafür gesorgt werden, dass bei der Iteration Sekundärsperrungen nur dann gesichert werden, wenn sie durch Untertabellenabgleich eingetragen wurden (vgl. 1.4.3). 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 nächsten Abschnitt gezeigt wird.

6.2 Rückführung von „überlappenden“ auf „vollständige“ Tabellen

6.2.1 Rückrechenbarkeit in sich sicherer und 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:

$$\begin{array}{lll} & X_1 + X_2 + X_3 & = 10 \\ & 0 + X_7 + X_8 & = 40 \\ \hline (1) & X_1 + (X_2 + X_3 + X_7 + X_8) & = 50 \end{array}$$

$$\begin{array}{lll} & X_2 + X_7 & = 20 \\ & X_3 + X_8 & = 15 \\ \hline (2) & (X_2 + X_3 + X_7 + X_8) & = 35 \end{array}$$

(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 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 solchermaßen 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

$a_1,a_2,a_3, \ldots$	Σ_1	...	$v_1, v_2, v_3,\ldots$	Σ_m	$\Sigma\,\Sigma$

Darin werden die Elemente der ersten Aggregationsstufe zu ihren Zwischensummen Σ_i zusammengefasst, die dann – ebenfalls aufaddiert – die Gesamtsumme $\Sigma\Sigma$ 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 zu einer Zwischensumme Σ^*_1 aufaddiert würden, dann die zweiten Elemente der ersten Aggretationsstufe zu Σ^*_2 usw., um dann die Zwischensummen Σ^*_j – die zweite Aggregationsstufe nach dieser neuen Gliederung – zur Gesamtsumme $\Sigma\,\Sigma$, der dritten Aggregationsstufe, zusammenzufassen. Dazu schreibt man zweckmäßig die 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,	a_2,	a_3, ...	Σ_1
b_1,	b_2,	b_3, ...	Σ_2
|	|	|	|
|	|	|	|
|	|	|	
v_1,	v_2,	v_3, ...	Σ_m
Σ^*_1,	Σ^*_2,	Σ^*_3, ...	$\Sigma\Sigma$

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 Σ_1, 15 b_i zur Summe Σ_2, usw. und nur 4 v_i zur letzten Zwischensumme Σ_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.

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 (Aggregat-)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 der Tabellendimension 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 einigen 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 sen-

sibler Daten verantwortlichen Fachstatistiker unter ausschließlicher Verwendung von Realdaten durchgeführt wird, oder wenn die zur Sicherung anstehenden Tabellendaten von vornherein, 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 demgemäß 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, das auszuschöpfen aber sehr rechenzeitaufwendig wäre; diese Idee wird lediglich in nachfolgendem Tabellenbeispiel angesprochen, bleibt aber sonst unberücksichtigt.

Im Folgenden wird die Behandlung von Dummies und Sternchensummen bei der Quaderauswahl für den Fall positiver Tabellen diskutiert. – Sind in der betrachteten Statistik positive und negative Tabellenwerte zu erwarten, kann Intervallschutz bisher nur bei konkreten vollständigen Tabellen mit Schätzintervallangaben ohne eingefügte Platzhalter gewährleistet werden (oben erwähnter Königsweg). –

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 aber nicht in der Quaderwertesumme erscheint (vgl. 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.

Bei Vorliegen positiver Tabellen 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 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 Dummy-Werte ausschließlich 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 genauso groß sind wie die 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 ursprünglichen unaufgestockten Tabelle eine Geheimhaltungslücke hinterließe. Dummies, die bezüglich irgendeiner Gliederung mit realen Tabellenwerten zu einer realen Randsumme beitragen, müssen wie strukturelle Nullen offen bleiben. Tragen sie zu einer Sternchensumme bei, so ist die Sternchensumme "erlaubtes" 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 Tabellendimension neu entstandenen Sternchensummen geeignet bewertet und gewichtet in den Eingabe-Datenbestand einzutragen:
   – 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 betragsmäßig 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 Dummys 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 Dummys 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-Zeit-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 (vgl. im zweiten Abschnitt Punkt 2.1.3).

   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 Punkt 7.1, graphische Darstellungen), bietet sich zumindest ein zweistufiges Vorgehen an, wonach im ersten Schritt alle primär geheimen Tabellenwerte auf unterstem Aggregationsniveau ohne Aufstokkung 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 am Schluss des Abschnitts 6.2.2.2 als „Technische Anmerkung".

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.

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

Abb. 6.4

Behebung der Rückrechenbarkeit der Beispieltabelle in Abb. 6.1

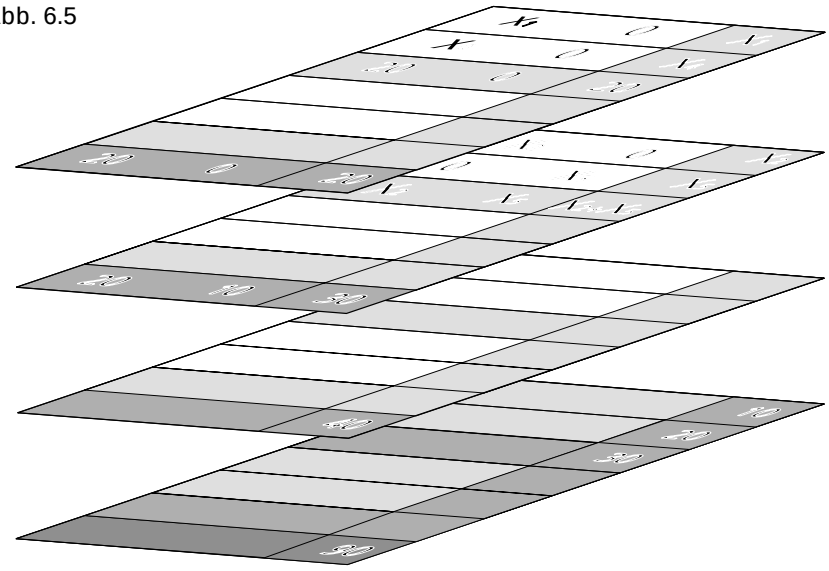
X_1	0	~~X_1~~	X_2	0	~~X_2~~	X_3	0	~~X_3~~	10 *
X_4	0	~~X_4~~	$0_\otimes$	X_5	~~X_5~~	0	X_6	~~X_6~~	20 *
~~20~~	~~0~~	~~20~~	~~X_2~~	~~X_5~~	~~X_2+X_5~~	~~X_3~~	~~X_6~~	~~X_3+X_6~~	30
0	0	0	X_7	0	~~X_7~~	X_8	0	~~X_8~~	40
0	0	0	0	X_9	~~X_9~~	0	X_{10}	~~X_{10}~~	20
~~0~~	~~0~~	~~0~~	~~X_7~~	~~X_9~~	~~X_7+X_9~~	~~X_8~~	~~X_{10}~~	~~X_8+X_{10}~~	60
20	0	20	20	10	30	15	25	40	90

$\otimes$ = 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.

Abb. 6.5



darüber liegenden Tabellenwerten bezüglich der neuen dritten Tabellendimension zu berechnen sind.

Die oberste zweidimensionale Tabelle dieser zur dreidimensionalen aufgestockten (aber noch nicht vollständigen) Tabelle (Abb. 6.5) 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: (siehe Abb. 6.6 auf S. 51).

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 darüberliegenden Summenzeilen zweiter Aggregationsstufe (siehe Abb. 6.6 auf S. 51).

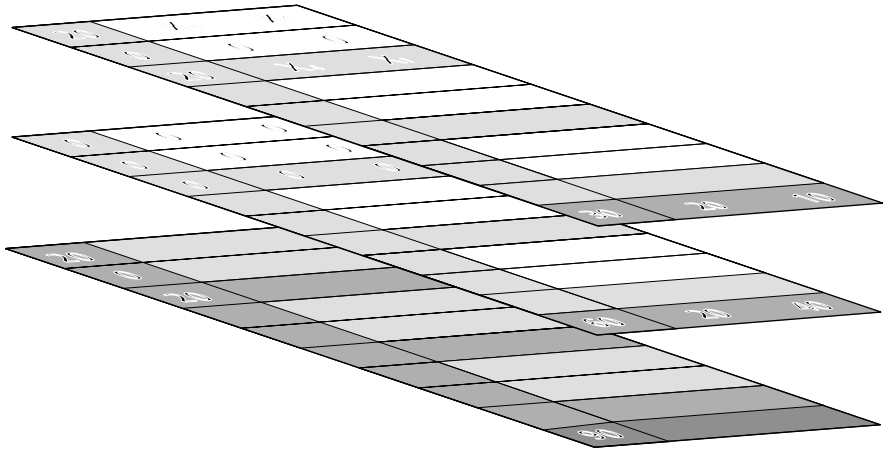
Die beiden oberen Streifen stimmen in ihren Sperrmustern bis auf die erste und dritte Spalte völlig überein. Aber auch die in der ersten und dritten Spalte des obersten Zeilenstreifens markierten geheimen Werte haben ihr „Gegenstück" und zwar im untersten Summenstreifen, weil ja Sternchensummen nicht 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 dazu die technische Anmerkung zum nachfolgenden Abschnitt).

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. in der aufgestockten Tabelle) dargestellt werden sollen. Die Antwort darauf gibt bereits das Quaderverfahren mit Untertabellenabgleich: Die neuen Gliederungsmerkmale sind in diesem Ver-

Abb. 6.6



fahren schon vorhanden; es sind die Positionsindizes, die die geometrische Lage der Untertabellen in der Gesamttabelle festlegen (vgl. Abschnitt 1.4.3 und insbesondere Abb. 1.8). Die Gesamtheit der Gliederungskriterien der aufgestockten Tabelle umfasst demnach die alten vorgegebenen Nutzerindizes mit den in den jeweiligen Gliederungen und Aggregationsstufen meisten Ausprägungen nebst Randsummenindizes (anschließende Indizes zur um 1 höheren Aggregationsstufe) und die Positionsindizes zu jedem alten Gliederungskriterium und zu jeder Aggregationsstufe samt den zugehörigen 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. Es liegt daher auch 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 Quadereckwerte dient. In dieser Arbeit wird aber 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 Anmerkung, 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 demgemäß 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 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

1. Schlüssel	2. Schlüssel: ACD	ACC	ACB	ACA	AC	ABC	ABB	ABA	AB	AAD	AAC	AAB	AAA	AA	A
0000134	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
0000133	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
0000132	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
0000131	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
0000130	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
0000125	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
0000124	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
0000123	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
0000122	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
0000121	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
0000120	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
0000113	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
0000112	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
0000111	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
0000110	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
0000100	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
Berichtspflichtige

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
0000134	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
0000133	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
0000132	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
0000131	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
0000131'	0 0 D	0 0 D	0 0 D	0 0 D	0 0 D	0 0 D	0 0 D	0 0 D	60 000 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
0000130	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
0000125	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
0000124	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
0000123	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
0000122	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
0000121	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
0000120	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
0000113	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
0000112	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
0000111	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
0000111'	0 0 D	0 0 D	0 0 D	0 0 D	0 0 D	0 0 D	0 0 D	0 0 D	60 000 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
0000111'	0 0 D	0 0 D	0 0 D	0 0 D	0 0 D	0 0 D	0 0 D	0 0 D	60 000 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
0000110	504 25 S	221 8 S	29 3	1 001 19	1 765 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
0000113'	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
0000112'	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
0000111'	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
0000111'	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
0000100	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 / Berichtspflichtige 100 P — Sperrvermerk (P = primär, S = sekundär, GD = gesperrte Dummies) — 10 000 / 100 SD (SD = sperrbare Dummies, D = nicht sperrbare Dummies)

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.

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 Eckwerte 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 zu ebenfalls realen Summenwerten aufaddiert werden können, sind gemäß Regel 2 als Eckwerte 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 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. Bei kleinen Tabellen bietet sich da 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~~ von oben gesehen vor 113 einordnete und die Zeile ~~131~~ vor 134. Dann lägen die Zeilen 112, 123 und 113, 134 mit Primärsperrungen im ersten bzw. zweiten Spaltenstreifen auf gleicher Zeilenposition bezüglich der Untertabellen und ließen sich so jeweils 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 demgemäß nur ein Karree von realen Tabellenwerten auf unterstem Aggregationsniveau ab. Die drei anderen Karrees mit den restlichen 12 geheimen Werten liegen als Projekti-

onen 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 Anmerkung“ 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 der technischen Anmerkung 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 ganz 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 eine Barriere 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 eine der bei Bearbeitung der vollständigen Tabelle erhaltenen Summensperrungen in der Zeile 130, 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 ist nicht zwingend auch 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 die oben beschriebene Sicherung mit zwei 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 vollständigen Tabelle 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 ln 60 000 / ln 61 000 * 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 = 456 750 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 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 im Allgemeinen 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.

Technische Anmerkung

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 Aufstokkung 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. Es bietet sich daher 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 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 unaufgestockten Tabelle als 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 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 parziell gesichert werden: Parzielle Sicherung bedeutet, dass in jeder Teiltabelle immer nur für diejenigen Primärsperrungen ein Quader mit der Dimension der betreffenden Teiltabelle aufzusuchen ist, die in dieser Teiltabelle die höchsten Aggregationsstufen aufweisen und deren Quader ganz 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 Dummies gründet: 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). Das 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 angewendet werden, die anderen „gewöhnlichen" Dummies und Sternchensummen werden weiterhin mit einheitlichen Gewichten und Werten belegt.

Schließlich kann man auf der Suche nach weiteren rechenzeit- und hauptspeicherplatzsparenden Aufstokkungsverfahren – wie am Ende von 6.2.2.1 bereits angedeutet – ganz auf die Erstellung einer aufgestockten Tabelle im Hauptspeicher oder auf anderen Datenträgern verzichten und statt dessen die vollständige Tabelle als Modellvorstellung nutzen, 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: Mit der Hilfskonstruktion „fiktive vollständige Tabelle" läßt sich für jeden primär geheimen Wert – ganz individuell – eine kleinste Teiltabelle so finden, dass zumindest ein Sicherungsquader ganz im Innern 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. Der mit solch einem Verfahren zu erreichende Rechenzeitgewinn wird durch Mehrfachbewertungen von Dummies zum Teil kompensiert (siehe 6.2.2.1). Dennoch sollte die „fiktive vollständige Tabelle" getestet werden – gerade auch im Hinblick auf den enormen Platzbedarf maschinenlesbar gespeicherter vollständiger Tabellen.

7. Anwendung des Quaderverfahrens auf Realdaten

7.1 Umsatzsteuerstatistik NRW 1994[9] als Beispiel für 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.

Die Umsatzsteuerstatistik ist mit ihren mehr als 700 000 Datensätzen eine im Vergleich zu anderen Statistiken für das Land Nordrhein-Westfalen sehr umfangreiche Tabelle, die besonders fein gegliedert ist. Die sehr feine Gliederung äußert sich in der sehr schwachen Besetzung mit zu weit über der 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.

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

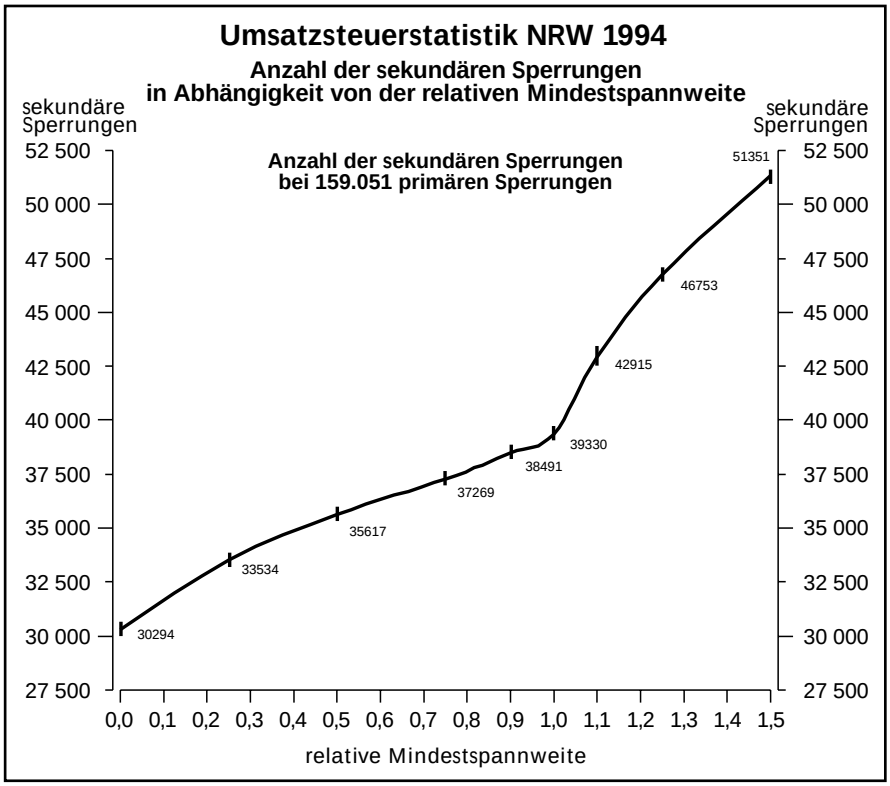
Informationen zur Sekundärsperrung

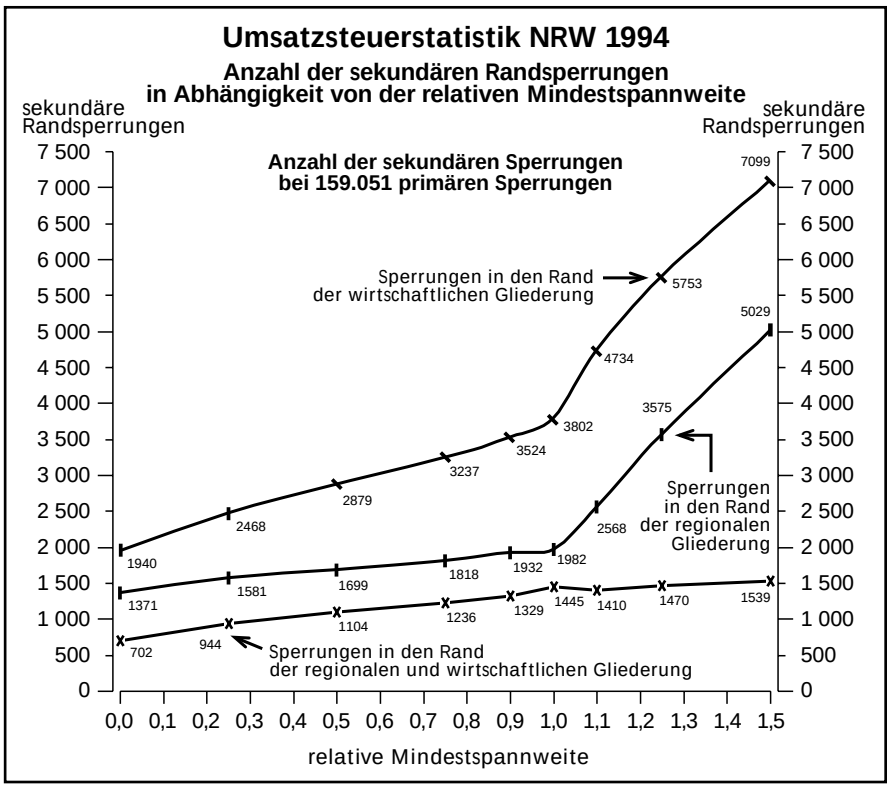
Rechenzeit (CPU-Zeit) bei Relativer Mindestspannweite gleich 0 (ohne Intervallschutz)	5min30
Rechenzeit (CPU-Zeit) bei Relativer Mindestspannweite gleich 1,5	5min52
Sekundäre Sperrungen bei – relativer Mindestspannweite gleich 0 – gesetzter Randschranke für beide Dimensionen[1)]	30 294
Sekundäre Sperrungen bei – relativer Mindestspannweite gleich 1,5 – gesetzter Randschranke für beide Dimensionen	51 351

1) Die für jede Dimension eingeführte Randschranke dient der Justierung, insbesondere bei überlappenden Tabellen: Bei erforderlichen Randsummensperrungen werden Summen mit Randschranke weitgehend gemieden.

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

9) Diese Auswertungen wurden auf IBM-9672 unter OS 390 mit der EDV-Programmversion GHQUAR.3 gemacht.





Randsummenwerte der betreffenden Untertabelle ausgewichen werden muss als bei Quaderauswahl ohne 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 graphischen Darstellungen (s. o.).

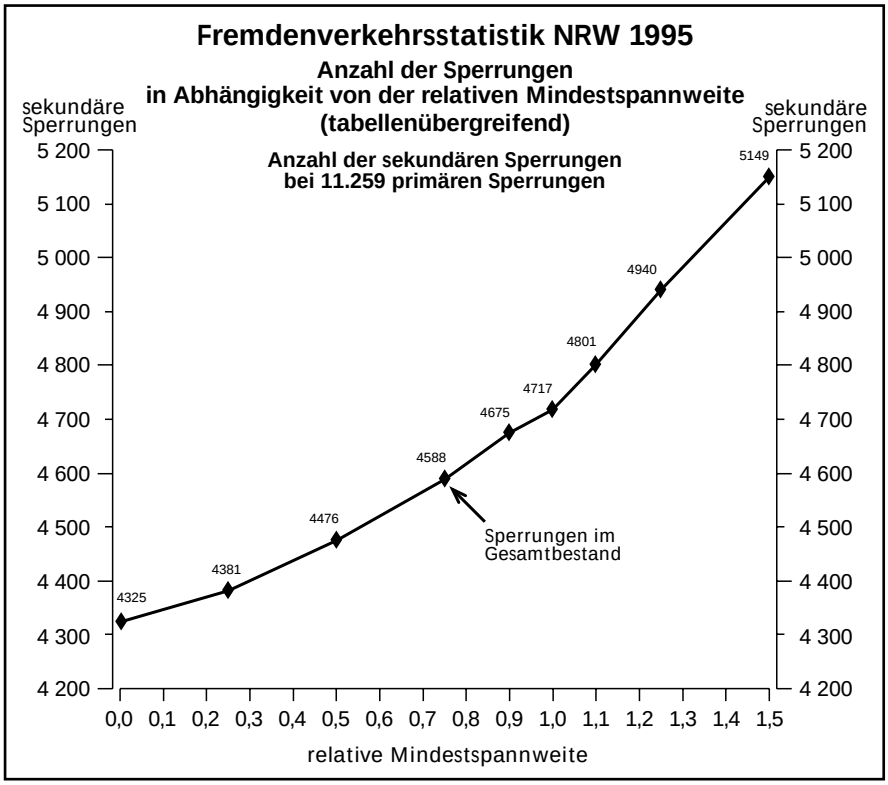
Die monoton mit der relativen Mindestspannweite zunehmenden Anzahlen von Sekundärsperrungen insgesamt (erste Umsatzsteuerstatistik-Darstellung) und in den Randsummen der Untertabellen (zweite Darstellung) weisen bei einer relativen Mindestspannweite von 1 einen Verlaufsknick auf (mit Ausnahme der Eckfeldsperrungen), 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. Demgemäß beobachtet man bei Gliederungen mit nur 2 Aggregationsstufen keinen so ausgeprägten Knick bei 100 % relative Mindestspannweite; die Fremdenverkehrsstatistik (Kap. 7.2) ist dafür ein Beispiel.

Dennoch bleibt die Gesamtzahl von Sekundärsperrungen auch im Falle der großen relativen Mindestspannweite von 150 % deutlich hinter der der 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.

7.2 Fremdenverkehrsstatistik NRW 1995[10] als Beispiel für 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 Ge-

10) Die Auswertungen wurden mit dem im LDS NRW entwickelten EDV-Programm GHMITER gemacht, das auf das durch GHQUAR.3 realisierte Quaderverfahren zurückgreift.



heimhaltung 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, die die Anzahl der Sekundärsperrungen in Abhängigkeit von der relativen Mindestspannweite wiedergibt.

Gesamttabelle:

Ausgangsdaten:

Datensätze (Tabellenfelder)	77 940
Primär geheime Werte	11 259
Leere Tabellenfelder	45 152

Informationen zur Sekundärsperrung

Sekundäre Sperrungen bei – relativer Mindestspannweite gleich 0 (ohne Intervallschutz) – gesetzter Randschranke für die 1. und 2. Dimension[10)]	4 325
Sekundäre Sperrungen bei – relativer Mindestspannweite gleich 1,5 – gesetzter Randschranke für die 1. und 2. Dimension	5 149
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. Einzeltabelle
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. Einzeltabelle
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 – relativer Mindestspannweite gleich 0 (ohne Intervallschutz) – gesetzter Randschranke für die 1. und 2. Dimension	2 211
Sekundäre Sperrungen bei – relativer Mindestspannweite gleich 1,5 – gesetzter Randschranke für die 1. und 2. Dimension	2 540

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

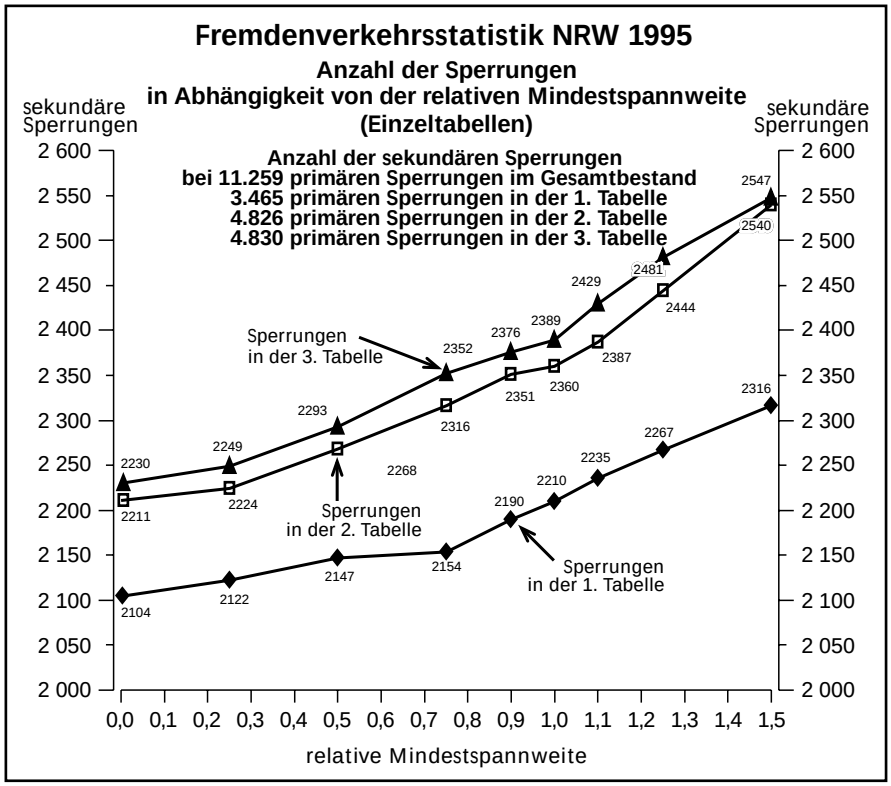
Informationen zur Sekundärsperrung

Sekundäre Sperrungen bei – relativer Mindestspannweite gleich 0 (ohne Intervallschutz) – gesetzter Randschranke für die 1. und 2. Dimension	2 230
Sekundäre Sperrungen bei – relativer Mindestspannweite gleich 1,5 – gesetzter Randschranke für die 1. und 2. Dimension	2 547

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. Demgemäß 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 Gesamt-Übersicht 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 Tabellenfelder 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 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.

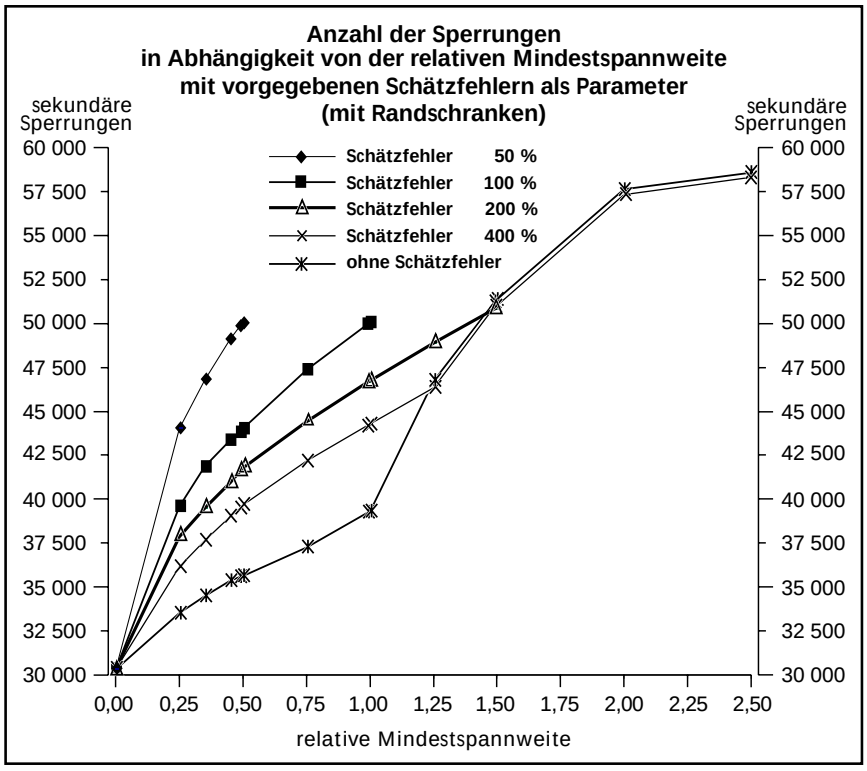
Dass die Freigabe der Tabellensummen für Sekundärsperrungen in der regionalen Gliederung zu einer wesentlichen Erhöhung der sekundären Sperreintragungen insgesamt führt, liegt in der Feinheit der Regionalstruktur begründet, die sich über vier Aggregationsstufen erstreckt. Die beiden anderen Gliederungen 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.

7.3 Berücksichtigung von externen Schätzintervallen am Beispiel der Umsatzsteuerstatistik NRW 1994

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 Sekundärsperrun-

gen des steuerbaren Umsatzes als positive Tabelle ohne Schätzintervalle zum Vergleich grafisch dargestellt.

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, nach denen hier die Quaderspannweite berechnet wurde, berücksichtigen diese Asymmetrie, indem sie das Schutzintervall als Schnittmenge aus dem symmetrischen, mit (9), Abschnitt 3.2 zu berechnenden Intervall und dem asymmetrischen Schutzintervall einer positiven Tabelle (gemäß (4), Abschnitt 3.1) bestimmen. 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.



Die Abbildung zeigt die Sekundärsperrungen in Abhängigkeit von der relativen Mindestspannweite mit vorgegebenem Schätzfehler als Parameter. Man sieht, dass sich der Kurvenverlauf mit zunehmendem Schätzfehler immer mehr abflacht, bis die Kurven bei ganz großen Schätzintervallen mit den Werten der Kurve der positiven Tabelle, bei der der Schätzfehler als beliebig groß angenommen werden kann, annähernd zusammenfallen, wenn sie nicht schon vorher abbrechen.

Die oben genannte Eigenschaft der Anzahl 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 7.1) –. Demgemäß nähern sich die Kurven mit zunehmendem Schätzintervallparameter auch immer mehr der Kurve der positiven Tabelle ohne Vorgabe von Schätzintervallen, bis sie mit dieser 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.

Die in der Abbildung gezeigte Beispieltabelle (siehe Seite 63) 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. 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

2. Schlüssel

1. Schlüssel	ACD	ACC	ACB	ACA	AC	ABC	ABB	ABA	AB	AAD	AAC	AAB	AAA	AA	A
0000134	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
0000133	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
0000132	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
0000131	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
0000130	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
0000125	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
0000124	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
0000123	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
0000122	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
0000121	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
0000120	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
0000113	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
0000112	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
0000111	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
0000110	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
0000100	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 / Berichtspflichtige

10 000 / 100 P — Sperrvermerk (P = primär, S = sekundär)

10 000 / 100 S2 — zusätzlicher Sperrvermerk (S2 = sekundär ohne Schätzfehler)

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.

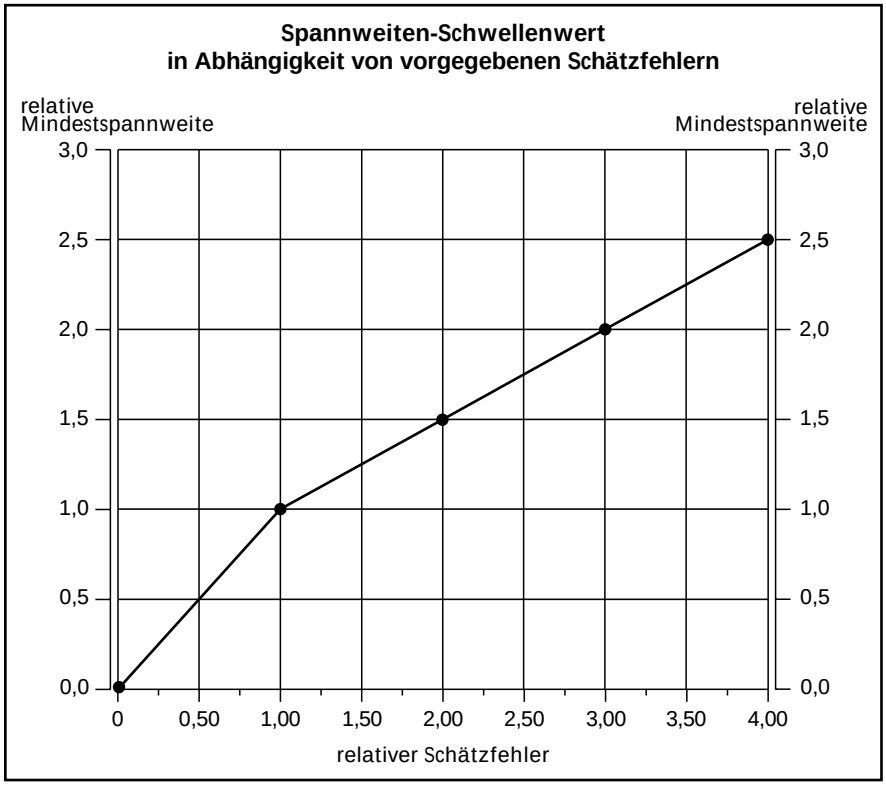
Um die Sekundärsperrungen in der Summenspalte AA im Falle der nur positiven Tabelle zu verstehen, betrachte man zunächst die Untertabelle der Spalten ACD, ACC, ACB, ACA und AC im obersten Tabellenstreifen. Das Einzelangabefeld (133; ACD) hat im Tabelleninneren der Untertabelle keinen Schutzquader, dessen auf seinen primär geheimen Wert bezogene Spannweite größer als die vorgegebene relative 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. Die Spannweite spielt bei der Sicherung sekundär geheimer Werte keine Rolle mehr!

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 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 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 eines entsprechenden Sicherungsquaders.

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 spielt, wie oben bemerkt, Intervallschutz keine Rolle. 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 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, 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 der selben 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 vor-ausgesetzt 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.

7.4 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 1.4.2 eingeführten Beispieltabelle vergleichbar; trotz ihrer starken Verdichtung

umfaßt sie aber immer noch mehr als zehnmal soviele 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 dem derzeit in der Anwendung befindlichen EDV-Programm GHQUAR ohne weiteres zu bearbeiten, da bei diesem Programm 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 (vgl. Absatz 2.1.3). Außerdem wurde vereinfachend auf Intervallschutz verzichtet. Wie vorhergehende Untersuchungen gezeigt haben (vgl. 7.1) 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 geheimgehalten werden müssen. In diesen Bereichen sind Primärsperrungen oftmals von vielen anderen primär geheimen Werten

Zur vollständigen 4-dimensionalen Tabelle der verkürzten Umsatzsteuerstatistik NRW 1994

Bestehend aus den Gliederungen:	regional -	kreisfreie Städte und Kreise Regierungsbezirke Land NRW
	wirtschaftlich -	Unterabschnitt Abschnitt Summe
Anzahl der gesamten/besetzten Tabellenfelder:		23 040 / 11 087
Anzahl der Dummies in den besetzten Tabellenfeldern		8535
Anzahl der Primärsperrungen/Einzelberichtspflichtigen		357 / 81
Anzahl der gesamten/besetzten Tabellenfelder (2-dim.)		*2 700 / 2 552*
Sicherung ohne Randschranke:		
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

Zur vollständigen 5-dimensionalen Tabelle der verkürzten Umsatzsteuerstatistik NRW 1994

Bestehend aus den Gliederungen :	regional –	kreisfreie Städte und Kreise Regierungsbezirke Land NRW
	wirtschaftlich –	Abteilung Unterabschnitt Abschnitt Summe
Anzahl der gesamten/besetzten Tabellenfelder:		138 240 / 119 346
Anzahl der Dummies in den besetzten Tabellenfeldern		113817
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 sekundär-Sperrungen	462
	CPU-Zeit in Sekunden	1 (normales Untertabellenverfahren)
5-dimensional:	Anzahl der sekundär-Sperrungen	464 (mit Sperrungen in den höchsten Aggrregationssstufen – Eckfeldsperrung)
	CPU-Zeit in Sekunden	750

umgeben, so dass sie sich gegenseitig schützen.

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 Aufstokkung 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.

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 Tabellendimension auf und sind infolgedessen 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 ausführlich hingewiesen, sie kam in dieser Arbeit jedoch nicht mehr zur Anwendung.

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.

8. Übersicht über Anwendungsmöglichkeiten des Quaderverfahrens

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 im Allgemeinen 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 Tabellenoraganisation 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 Kapitel-Nummern, unter denen diese Faktoren abgehandelt werden, und Bemerkungen mit anwendungsrelevanten Hinweisen.

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 Gesamtda-

Quaderverfahren für n-dimensionale Tabellen					
Faktor	Faktorkategorie/Kapitel/Bemerkung				
Tabellentyp	Wertetabelle ohne Zwischen-summen	Wertetabelle mit Zwischen-summen	überlappende Tabellen	Zeitreihen-Tabellen	Kontingenz-tabellen
	1.1 bis 1.3, 2. u. 3., (6.2.2)	1.4	6.1	5.3.2	Einführung
	ohne Zwischen-summen gegeben oder Aufstockung	Untertabellen-hierarchie	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 oder Intervallschutz	zu kleine Schätzintervalle verhindern Geheimhaltung		
Tabellen-organisation	Untertabellen-Hierarchie mit Abgleich	Vervollständigung durch Aufstockung	Zusammenführung überlappender Tabellen in gemeinsamen Datenbeständen		
	1.4	6.2	6.1		
	Verfeinerung zu Tabellenteilen mit Rand- ohne Zwischensummen	Vergröberung zu Gesamttabelle ohne Zwischensummen	allen Tabellen gemeinsamen Aggregate werden nur einmal aufgeführt		
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 eines Quaders	gerade indiziert: $+\varepsilon$ ungerade indiziert: $-\varepsilon$	
Grad der Sicherung	nur notwendiger, kein hinreichender Schutz bei Tabellenabgleich		hinreichender Schutz nur bei Tabellen ohne Zwischensummen		
	6.2.1		6.2		
	Beispiele: Untertabellenabgleich, Abgleich von überlappenden Tabellen		zu erreichen durch Vermeidung von Summensperrungen und durch Dimensionsaufstockung		

ten. 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 Aufstokkung 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 Summen-Sekundärsperrungen insbesondere in höhere 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 Tabellen mit iterativem Abgleich gesichert werden können – hierzu gibt es ebenfalls Gegenbeispiele -. 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!

   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 über mehrere Zeitperioden).

Literaturangaben:

Appel, Günther / Kinzel, S. / Nölte, Dieter, Statistisches Landesamt Berlin, 1992: SAFE – A Generally Usable Program System for the Anonymization of Individual Data in Official Statistics. Vortrag beim International Seminar on Statistical Confidentiality in Dublin

Cox, Lawrence H., U.S. Bureau of the Census: Suppression Methodology and Statistical Disclosure Control, in: Journal of the American Statistical Association, 1980, Vol. 75, No 370

Cox, Lawrence H., U.S. Bureau of the Census: Linear Sensitivity Measures and Statistical Disclosure Controll, in: Journal of Statistical Planning and Inference, 1981, 5, S. 153 - 164

Cox, Lawrence H., 1992: Solving Confidentiality Protection Problems in Tabulations using Network Optimization: A Network Model for Cell Suppression in U.S. Economic Censuses. Vortrag beim International Seminar on Statistical Confidentiality in Dublin

Geurts, Jacqueline, Netherlands Central Bureau of Statistics, Department of Statistical Methods, P.O. Box 959, 2270 AZ Voorburg, The Netherlands, 1992: Heuristics for Cell Suppression in Tables

Repsilber, Rüdiger D.: EDV-Verfahren zur Wahrung der Geheimhaltung bei Tabellen mit bis zu sieben Ordnungskriterien, in: Statistische Rundschau Nordrhein-Westfalen Februar 1991, Landesamt für Datenverarbeitung und Statistik Nordrhein-Westfalen, Düsseldorf 1991, S. 78 - 84

Repsilber, Rüdiger D., Landesamt für Datenverarbeitung und Statistik Nordrhein-Westfalen, 1992: Safeguarding Secrecy in Aggregative Data. Vortrag beim International Seminar on Statistical Confidentiality in Dublin

Repsilber, Rüdiger D., Landesamt für Datenverarbeitung und Statistik Nordrhein-Westfalen, 1994: Preservation of Confidentiality in Aggregated Data. Vortrag beim International Seminar on Statistical Confidentiality in Luxembourg

Robertson, Dale A., Statistics Canada, 1994: Automated Disclosure Control at Statistics Canada. Vortrag beim International Seminar on Statistical Confidentiality in Luxembourg

Zayatz, Laura V., U.S. Bureau of the Census, 1992: Using Linear Programming Methodology for Disclosure Avoidance Purposes. Vortrag beim International Seminar on Statistical Confidentiality in Dublin