Technische Universität Ilmenau
Fakultät für Wirtschaftswissenschaften
Institut für Wirtschaftsinformatik
Fachgebiet für Wirtschaftsstatistik und Operation Research
Univ.-Prof. Dr. Martin Boeselt

Beschreibung von Verfahren zur statistischen Geheimhaltung in Tabellen und ihre Anwendung

Freie wissenschaftliche Arbeit
zur Erlangung des akademischen Grades
Diplom Wirtschaftsingenieur

an der Fakultät für Wirtschaftswissenschaften
der Technischen Universität Ilmenau

Referent:	Prof. Dr. M. Boeselt
Betreuer:	Prof. Dr. K. Luhn
Bearbeiter:	Pierre Wettig
Bearbeitungszeit:	6 Monate
Abgabetermin:	08. Februar 2002

Inhaltsverzeichnis

1. Einleitung

1.1. Gegenstand der Arbeit

Die meisten Menschen wissen, dass die Bedeutung der elektronischen Datenverarbeitung zunimmt. Doch zu wenige haben erkannt, dass damit Daten und Fakten auch immer wichtiger werden. Unsere Industriegesellschaft ist komplex, und dennoch soll sie reibungslos funktionieren:

- Arbeitnehmer und Aktionäre erwarten, dass die Manager sich gut informieren und rechtzeitig die richtigen Entscheidungen treffen,
- Bürger wollen, dass Politiker effektiv wirtschaften, Entwicklungen rechtzeitig erkennen und sorgfältig planen,
- Wissenschaftler benötigen für Analysen oder Untersuchungen objektive Zahlen.

Die amtliche Statistik liefert die Datenbasis für die Entscheidungsträger, mit der sie ihre Betriebe auch durch krisengeschüttelte Zeiten steuern können. Ohne Statistiken könnte unser Sozialstaat nicht planen und Bilanz ziehen.
Die Statistikämter können nur dann sinnvolle Gesamtzahlen bereitstellen, wenn alle mitwirken. Deshalb hat der Gesetzgeber die Auskunftspflicht festgelegt. Im Übrigen werden die Angaben streng vertraulich behandelt. Die gesetzliche Geheimhaltungspflicht gilt auch gegenüber anderen Behörden, wie beispielsweise den Finanzämtern. Statistische Ergebnisse sind anonym. Die Daten werden nur zusammengefasst veröffentlicht und somit sind keine Einzelfälle erkennbar.

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.

Bei der mathematischen Untersuchung von Offenlegungsmethoden lassen sich zwei Konzepte unterscheiden. Zum einen geht es um Verfahren zur Verhinderung der Offenlegung von tabellierten Daten (Makrodaten). In der Praxis wird dieses Konzept seit jeher meist im Zusammenhang mit der Verbreitung von Daten aus Industrie und Wirtschaft angewandt. Das andere Konzept ist auf Mikrodateien mit personenbezogenen Daten (Mikrodaten) ausgerichtet und betrifft Verfahren zur Verringerung der Offenlegungsmöglichkeiten durch externe Benutzer solcher Dateien. Dieses Konzept wird in erster Linie bei Zählungen und Erhebungen über die Sozialstruktur angewandt. In beiden Fällen wird hauptsächlich die Verteilung und Vermeidung von seltenen Elementen der Grundgesamtheit untersucht, und zwar in Feldern von hochdimensionierten Tabellen oder als praktisch eindeutige Phänomene in Grundgesamtheiten und Stichproben. Zur Schätzung der Eindeutigkeitsrate in einzelnen Dateien oder mehrdimensionalen Tabellen werden verschiedene Verteilungen herangezogen. Im Mittelpunkt dieser Diplomarbeit stehen verteilungsfreie Schätzverfahren. Ziel dieser Verfahren ist nicht die genaue Schätzung des Offenlegungsrisikos, sondern die Berechnung der damit verbundenen Parameter, z.B. die Verteilung der Anzahl der Felder in einer Tabelle mit 0, 1, 2,..., k Elementen. Es existiert eine Vielzahl von Transformationsverfahren, auf die im weiteren noch konkret eingegangen wird. Beispielhaft sollen an dieser Stelle die Unterdrückung, die Umkodierung und die stochastische Perturbation genannt werden.

Klassifizierung statistischer Daten[1]

a) Mikrostatistiken beinhalten ein Set von individuellen Daten, z.B. Personen, Haushalte, Unternehmen. Zu jedem Subjekt gehört ein individueller Datenvektor, welcher qualitative und/oder quantitative Attribute enthalten kann. Das folgende Beispiel soll zeigen, wie Mikrodaten in Form einer Tabelle aufbereitet werden können:

Alter	Geschlecht	ehelicher Status	Anzahl der Kinder	Einkommen	...
41	männlich	ledig	0	70000	...
38	weiblich	verheiratet	3	95000	...
65	weiblich	verwitwet	2	30000	...
...	...	...	...	...	...

b) Makrostatistiken sind Tabellen, die aggregierte Einzelangaben beinhalten. Das folgende Beispiel zeigt eine solche Tabelle:

	Gruppe A	Gruppe B	Gruppe C	Total
Männer	10	1	39	50
Frauen	30	18	2	50
Total	40	19	41	100

Gegenstand dieser Arbeit sind aggregierte Daten in Tabellen, d.h. ausschließlich der Umgang mit makrostatistischen Daten.

1.2. Zielsetzung

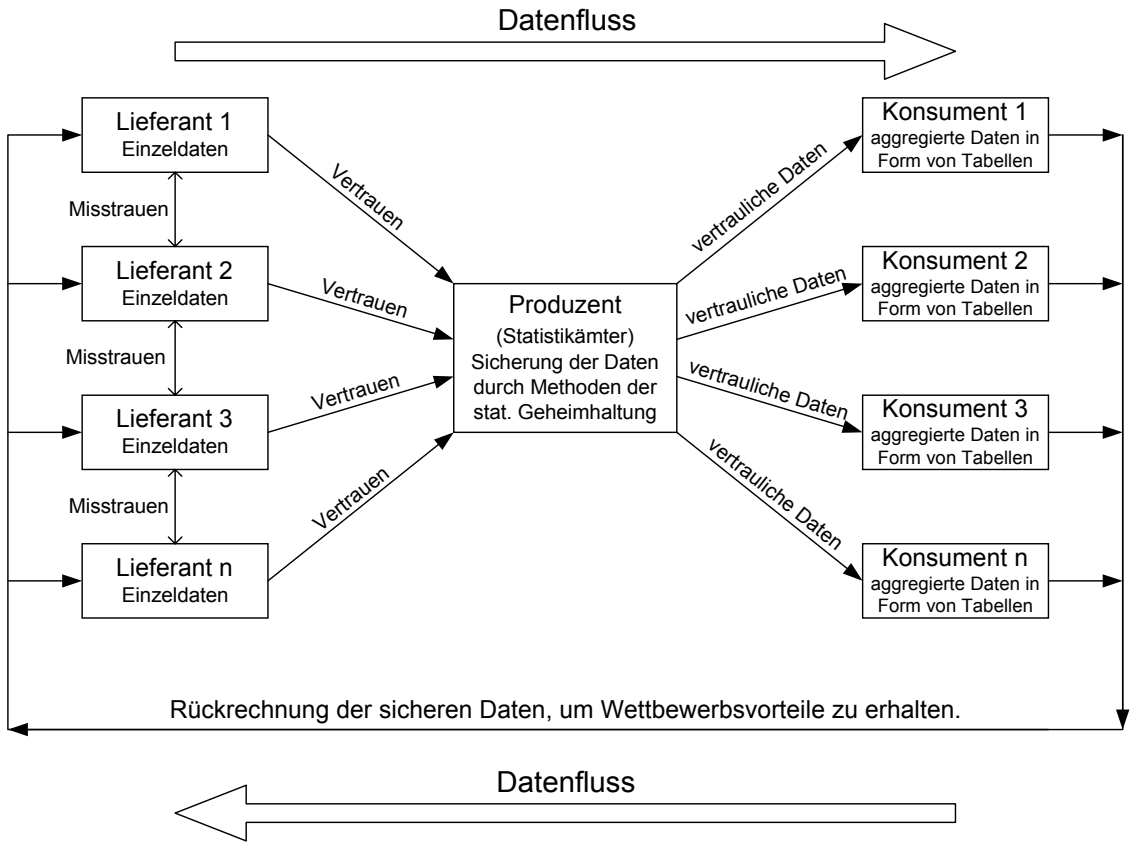
Ziel dieser Arbeit ist die Zusammenstellung, Katalogisierung und Bewertung aller derzeit verwendeten Geheimhaltungsverfahren. Dabei wird dieses Problem stets aus zwei Perspektiven betrachtet. Es stehen sich die Gruppe der Produzenten (meist Statistikämter) und die Gruppe der Konsumenten (Benutzer der aufgearbeiteten Tabellen) gegenüber. Die Lieferanten von Daten sind auch gleichzeitig die Benutzer der aggregierten Tabellen. Dieser Zusammenhang zwischen Lieferant und Konsument liefert Konfliktpotential. Einerseits möchte

[1] Dalenius, T; 1977

jeder „Datenlieferant“ seine Angaben unter Datenschutzgesichtspunkten gesichert wissen, andererseits strebt er selbst auch nach vollkommener

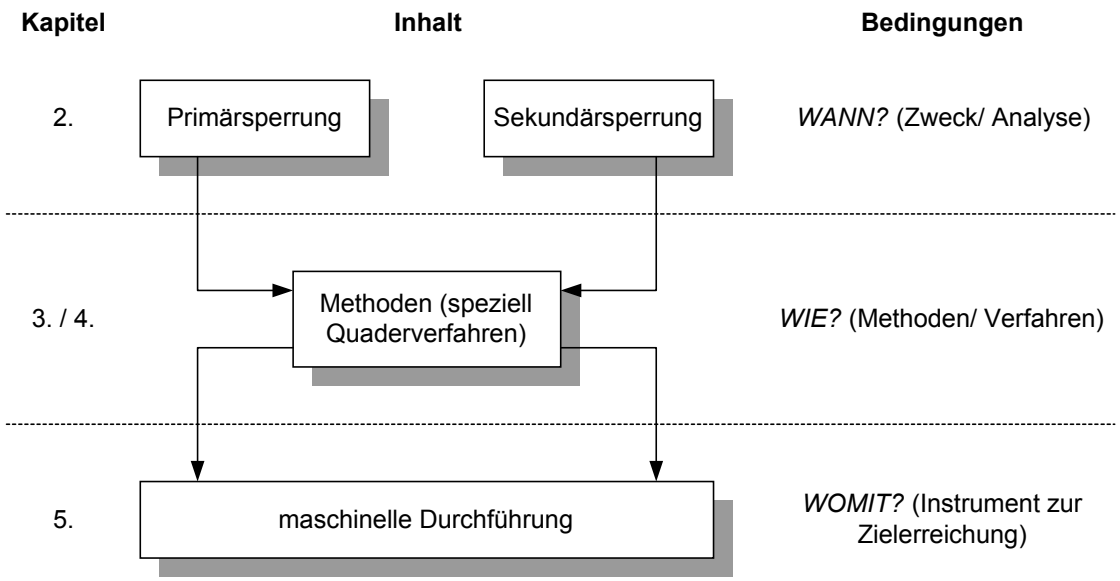
Information. Die Aufgabe der Statistikämter besteht nun darin, mit geeigneten Methoden das Offenlegungsrisiko von Einzelangaben zu minimieren, gleichzeitig aber den Informationsgehalt veröffentlichter Tabellen so hoch wie nur möglich zu halten.

Die Wertschöpfungskette (Entstehung und Versuch der Aufdeckung gesicherter Tabellen) kann wie folgt veranschaulicht werden:



1.3. Aufbau der Arbeit

Diese Arbeit ist in vier wesentliche Hauptgruppen eingeteilt. Nach dem einleitenden Kapitel 1, beschäftigt sich Kapitel 2 mit der Analyse der zu verarbeitenden Daten. Hierbei werden sensible Tabellenfelder lokalisiert (Primär- und Sekundärsperrung). Es werden Bedingungen entwickelt, unter denen bestimmte Daten als sensibel eingestuft werden müssen. In Kapitel 3 werden Methoden und Verfahren vorgestellt, mit denen der eigentliche Sicherungsprozess durchgeführt werden kann. Anhand kleiner, übersichtlicher Tabellen soll gezeigt werden, wie vielfältig die Möglichkeiten zur Sicherung sensibler Tabellenfelder sind. Diese Verfahren und Analysen in Kapitel 2 und 3 bilden die Grundlage für die maschinelle Durchführung (Kapitel 5) des Sicherungsprozesses. Aufgrund des unüberschaubaren Umfangs dieser Tabellen ist eine Sicherung von „Hand“ unmöglich. Man muss davon ausgehen, dass diese Tabellen mehrere 100000 Einträge in verschiedenen Hierarchiestufen besitzen. Das vom Statistischen Landesamt NRW entwickelte Quaderverfahren wird explizit in Kapitel 4 vorgestellt.



2. Statistische Geheimhaltung in Tabellen

2.1. Die Primärsperrung

Gemäß §16 (europäisches Recht auf Grundlage der Verordnung Nr. 1588/90 des Rates) des Bundesstatistikgesetzes unterliegen Statistiken der Geheimhaltungspflicht von Einzelangaben. Ausnahme hierbei bildet §16 Abs. (1), 3. und 4., der besagt, dass dies nicht für Einzelangaben gilt, die dem Betroffenen nicht direkt zuzuordnen sind oder für aggregierte Daten (d.h. mehrere Einzelangaben werden beispielsweise in Tabellen zusammengefasst). Des weiteren verzichtet das Statistische Bundesamt auf Veröffentlichungen aggregierter Daten, wenn dabei die Gefahr einer exakten Aufdeckung einzelner Daten möglich ist[2]. Dies kann durch verschiedene Rückrechnungsalgorithmen möglich sein, wenn Tabellen nicht ausreichend gegen Angriffe geschützt werden.
Wie eine solche Tabelle gegen exakte und/oder gegen näherungsweise Offenlegung geschützt werden kann, wird im Folgenden aufgezeigt.
Die Tabellenfelder, bei denen die Gefahr einer exakten bzw. näherungsweisen Offenlegung besteht, müssen geheimgehalten werden. Die Unterdrückung dieser Tabellenfelder wird als primäre Geheimhaltung bezeichnet. Neben der primären Sperrung von Tabellenfeldern müssen noch zusätzliche Werte sekundär unterdrückt werden, um eine Aufdeckung der primär geschützten Zellen zu verhindern. Die Sekundärsperrung wird meist nicht aus Datenschutzgründen durchgeführt, sondern zur Erhöhung der Datensicherheit von sensiblen (primär) Tabellenfelder.

Mit Hilfe verschiedener Geheimhaltungsregeln werden die primär zu sperrenden Werte ermittelt. Zwei Kriterien finden dabei Beachtung:

1. Welches Vorwissen wird einem Angreifer unterstellt?
2. Führt nur eine mögliche exakte Offenlegung zur Sperrung der Zelle oder genügt eine näherungsweise Ermittlung der geheimzuhaltenden Werte?

[2] DA-Teil A Nr.1 zu §75 der Geschäftsordnung des Statistischen Bundesamtes

2.1.1. Vermeidung einer exakten Offenlegung[3]

Diese Regeln, welche eine exakte Offenlegung verhindern, werden auch als Fallzahlregeln bezeichnet. Hierbei wird ein Wert geheimgehalten, wenn eine bestimmte Mindestfallzahl, der zu diesem Tabellenfeld beitragenden Einzelangaben unterschritten wird. Die Anzahl der zum Wert einer Zelle beitragenden Einzeldaten werden mit n bezeichnet. Folgende Szenarien sind nun denkbar:

- n=1: Die Geheimhaltung ist nicht gesichert, da es sich nicht um einen aggregierten Zellenwert handelt, sondern um eine Einzelangabe, weil der Tabellenwert nur aus einer Einheit besteht

- n=2: Die Geheimhaltung ist nicht gesichert, weil jeder der beiden Befragten mit seinem Vorwissen (d.h. sein eigener Beitrag zum Zellwert) durch einfache Differenzbildung den Beitrag des Anderen am Gesamtwert der Zelle exakt ermitteln kann

- n=3: Die Geheimhaltung ist nur unter der Bedingung gesichert, dass jeder Angreifer nur einen Einzelbeitrag (z.B. seinen eigenen) kennt. Da man jedoch nicht von einem beschränkten Vorwissen eines Angreifers ausgehen kann, wird es auch hier keinen vollkommenen Schutz geben. Durch Absprachen zweier Einheiten zum Nachteil des Dritten, könnten die Einzelangaben aufgedeckt werden. Das Kriterium „Vorwissen“ eines Angreifers ist also nur schwer kalkulierbar.

- Sofern also mehr Vorwissen über Einzelbeiträge unterstellt werden kann, sind Fallzahlen mit n>3 sinnvoll. Wenn anzunehmen ist, dass ein Angreifer n-1 Einzelbeiträge kennt, so ist für dieses Tabellenfeld eine Mindestfallzahl von n+1 Einzelangaben erforderlich, um eine exakte Aufdeckung zu verhindern.

[3] Gießing, S.; 1999

2.1.2. Vermeidung einer näherungsweisen Offenlegung

Wenn eine näherungsweise Offenlegung von Einzelangaben verhindert werden soll, reicht die Anwendung der Fallzahlenregeln nicht mehr aus. Eine näherungsweise Offenlegung von Einzelangaben ist dann möglich, wenn ein Tabellenwert von einer Einzelangabe dominiert wird. Zur Sicherung dieser Zelle kommen die sogenannten (n,k)-Dominanzregeln zur Anwendung.

2.1.2.1. Die (1,k) - Dominanzregel

Diese Regel besagt, dass der Wert X eines Tabellenfeldes geheimzuhalten ist, wenn der größte Einzelbeitrag x_1 mehr als k% des Zellwertes beträgt, d.h. wenn gilt[4]:

$$\boxed{x_1 > \frac{k}{100} \cdot X} \qquad \text{(A)}$$

Ein Angreifer, der den Wert des größten Einzelbeitrages gleich dem Gesamtzellwert schätzt, wird den wahren Wert x_1 um mindestens $100 \cdot \frac{100-k}{k}\,\%$ verfehlen.

Dies ergibt sich aus folgender Äquivalenzumformung:

$$x_1 \le \frac{k}{100} \cdot X \quad \leftrightarrow \quad \frac{\hat{x}_1}{x_1} \ge \frac{100}{k} \quad \leftrightarrow \quad \frac{\hat{x}_1 - x_1}{x_1} \cdot 100 \ge \left(\frac{100}{k} - 1\right) \cdot 100$$

Im Statistischen Bundesamt wird im Allgemeinen die (1,85)-Dominanzregel angewandt. Das heißt, Tabellenfelder müssen geschützt werden, wenn der größte Einzelbeitrag zum Gesamtwert der Zelle 85% übersteigt.

[4] Cox, L.H.; 1981

Hierzu folgendes Beispiel 1[5]: Es gilt die (1,85)-Dominanzregel

- Der Gesamtwert des Tabellenfeldes betrage X=1000 €
- Der Wert des größten Einzelbeitrages sei x_1=851 €.

Nach Formel A: 851 € $> \left(\frac{85}{100}\right) \cdot 1000$ € wäre dieses Tabellenfeld mit dem Wert

X=1000 € zu sperren, weil es von einem Einzelbeitrag mit mehr als 85% dominiert wird.

Angenommen x_1 betrage 850 €, dann bräuchte die Zelle nicht geheimgehalten zu werden. Würde nun ein Angreifer den größten Einzelwert x_1 auf 1000 € schätzen, so würde er den wahren Wert von x_1 um 17,6% verfehlen. Eine wesentlich genauere Schätzung wird genau dann möglich, wenn ein Angreifer sein Vorwissen mit einbezieht. Der Befragte mit dem zweitgrößten Einzelbeitrag kann, indem er seinen eigenen Wert x_2 vom Zellwert abzieht (x_1=X-x_2), den größten Einzelwert wesentlich genauer schätzen, wie folgendes Beispiel 2 verdeutlicht.

Es gelten die Annahmen aus Beispiel 1. Der Einzelwert des Angreifers beträgt x_2=120 €, der größte Einzelbeitrag x_1=850 €. Nach Beispiel 1 muss das Tabellenfeld mit dem Wert X=1000 € nach der (1,85)-Dominanzregel nicht geheimgehalten werden.

Der Angreifer schätzt x_1 auf 880€ (x_1=1000 € – 120 €).

Nun gilt: $100 \cdot \frac{880 - 850}{850} \approx 3{,}5$

Das heißt, der Wert des größten Einzelbeitrages wird vom Angreifer mit dem Vorwissen über seinen eigenen Beitrag (x_2) um nur noch 3,5% überschätzt. Ohne Vorwissen waren es noch 17,6%. Der Angreifer kann den größten Einzelbeitrag also nahezu aufdecken. Dieses Beispiel zeigt, dass analog zur Fallzahlregel n=2, die Geheimhaltung in bestimmten Fällen nicht gesichert ist.

[5] Beispiele in diesem Kapitel aus: Gießing, S.; Forum der Bundesstatistik (teilweise bedarfsgerecht verändert)

2.1.2.2. Die (2,k) - Dominanzregel

Diese Regel besagt, dass der Tabellenwert X geheimzuhalten ist, wenn er mit mehr als k% durch die Summe der beiden größten Einzelangaben dominiert wird, d.h. wenn gilt[6]:

$$\boxed{x_1 + x_2 > \frac{k}{100} \cdot X} \qquad \text{(B)}$$

Nach der (2,85)-Dominanzregel wäre der Tabellenwert aus Beispiel 2 X=1000 € geheimzuhalten, weil x_1+x_2=970 € 85% des Gesamttabellenwertes übersteigt.
Das heißt, dass Einzelangaben die den Anforderungen der (2,85)- Dominanzregel genügen, wesentlich besser geschützt sind. Diese Regel ist äquivalent zu der Fallzahlenregel n=3. Es wird also unterstellt, dass jeder Angreifer nur seinen eigenen Beitrag zum Gesamtwert der Zelle kennt.

2.1.2.3. Die (n,k) - Dominanzregel

Wenn unterstellt werden muss, dass Vorwissen über mehr als einen Einzelbeitrag besteht, kann die (n,k)-Dominanzregel mit n>2 sinnvoll sein. Nach dieser Regel ist ein Tabellenfeld dann geheimzuhalten, wenn die n größten Einzelbeiträge zusammen den Gesamtwert der Zelle um mehr als k% dominieren. Ist anzunehmen, dass ein Angreifer über Vorwissen von n-1 Einzelbeiträgen verfügt, so ist für dieses Tabellenfeld die (n,k)-Dominanzregel erforderlich, um eine näherungsweise Aufdeckung eines Einzelbeitrages zu verhindern[7].

2.1.2.4. Die (p%) – Regel

Nach dieser Regel ist ein Tabellenfeld dann geheimzuhalten, wenn die Differenz zwischen Gesamtzellwert und dem zweitgrößten Einzelwert, den größten Einzelwert um weniger als p% übersteigt[8],

[6] Willenborg, L./ de Waal, T; 1998
[7] Willenborg und de Waal; 1998
[8] Kelly, Golden und Assad; 1990

d.h. wenn:

$$\boxed{\frac{(X - x_2) - x_1}{x_1} \cdot 100 < p} \qquad \text{(C)}$$

Der Angreifer mit dem Vorwissen über den Wert des zweitgrößten Einzelbeitrages, soll also den größten Einzelbeitrag um mindestens p% überschätzen. Der Schwellenwert für p ist vorab festzulegen. Er sollte jedoch nach Beispiel 1 eine Mindestüberschätzung von p≥17,6% gewährleisten. Es werden dann also nur Tabellenwerte veröffentlicht, bei denen die Schätzung des dominierenden größten Einzelbeitrages den wahren Wert um mindestens 17,6% verfehlt.

Beispiel 3:
Es gelte die 17,6%-Regel → p=17,6

- Wert des Tabellenfeldes X=1000 €
- Wert des größten Einzelbeitrags x_1=850 €
- Wert des zweitgrößten Einzelbeitrags x_2=120 €

Damit gilt: $\hat{x}_1 = X - x_2 = 1000€ - 120€ = 880€$

$$\rightarrow \; 100 \cdot \frac{\hat{x}_1 - x_1}{x_1} = 100 \cdot \frac{880€ - 850€}{850€} \approx 3{,}5$$

d.h. der Wert des größten Einzelbeitrages kann vom Angreifer mit dem Vorwissen des zweitgrößten Einzelbeitrages auf 3,5% genau geschätzt werden. Damit wäre diese Zelle nach der 17,6%-Regel geheimzuhalten. Die Summe der zwei größten Einzelwerte beträgt: x_1+x_2 =850 €+120 € =870 €.

Weil 870 € $> \left(\frac{85}{100}\right) \cdot 1000$ €, würde diese Zelle auch nach der (2,85)- Dominanzregel gesperrt werden. Die 17,6%-Regel ist allerdings der (2,85)-Dominanzregel überlegen. In bestimmten Fällen müssen aufgrund der (2,85)-Dominanzregel Tabellenfelder gesperrt werden, die jedoch bei Anwendung der 17,6%-Regel veröffentlicht werden können. Das ist ein Vorteil, weil bei der Optimierung solcher Aufgaben ein möglichst minimaler Informationsverlust angestrebt wird, d.h. es sollen nur so viele Zellen gesperrt werden, wie unbedingt nötig.

Beispiel 4:

Es gelte die 17,6%-Regel → p= 17,6

- Wert des Tabellenfeldes X=1000 €
- Wert des größten Einzelbeitrags x_1=500 €
- Wert des zweitgrößten Einzelbeitrags x_2=400 €

Damit gilt nach Formel C: $\hat{x}_1 = X - x_2 = 1000€ - 400€ = 600€$

$$\rightarrow \quad 100 \cdot \frac{\hat{x}_1 - x_1}{x_1} = 100 \cdot \frac{600€ - 500€}{500€} = 20 > 17{,}6$$

d.h. der Wert des größten Einzelbeitrags kann von dem Angreifer mit dem Vorwissen über den zweitgrößten Einzelbeitrag nur auf 20% genau geschätzt werden und dürfte nach der 17,6%-Regel veröffentlicht werden. Die Summe der zwei größten Einzelwerte beträgt aber x_1+x_2 =900 €.

Weil 900 € > $\left(\frac{85}{100}\right) \cdot 1000$ €, muss das Tabellenfeld nach der (2,85)- Dominanzregel geheimgehalten werden.

2.1.2.5. Die (p;q) - Regel

Ist zu unterstellen, dass ein Angreifer über Vorwissen von mehr als einem Einzelbeitrag verfügt, dann kommt die sogenannte (p;q)-Regel zum Einsatz. Da man jedoch nie genau abschätzen kann, welches Vorwissen ein Angreifer besitzt, wird es immer sinnvoll sein, die (p;q)-Regel anzuwenden. Hierbei kann man von einem wesentlich höheren Schutzniveau als bei den anderen Regeln ausgehen. Ein Angreifer kann jeden Einzelbeitrag auf $\pm$ q% genau schätzen.
Nach der (p;q)-Regel sind Tabellenfelder geheimzuhalten, wenn es dem Befragten mit dem zweitgrößten Einzelbeitrag möglich ist, den größten Einzelbeitrag auf $\pm$ p% genau zu schätzen, indem er seinen eigenen Beitrag, sowie die von ihm aufgrund seines Vorwissens auf q% genau geschätzten übrigen Einzelbeiträge vom Gesamtzellenwert abzieht,

d.h. wenn gilt:

$$\boxed{X - x_2 - \sum_{i=3}^{N} \hat{x}_i - x_1 < \frac{p}{100} \cdot x_1} \qquad \text{(D)}$$

Hierbei entspricht $X - x_2 - x_1$ den restlichen Einzelbeiträgen des Zellwertes. Nach der Annahme weicht der Schätzwert der restlichen Beiträge um höchstens q% vom wahren Wert des Restes ab.

$$\left(\sum_{i=3}^{N} x_i - \sum_{i=3}^{N} \hat{x}_i \leq \frac{q}{100} \cdot \sum_{i=3}^{N} x_i \right)$$

Das bedeutet, dass ein Tabellenfeld nach der (p;q)-Regel geheimgehalten wird, wenn q% der Restwerte kleiner als p% des größten Einzelwerts ist,

d.h. wenn gilt: $\frac{p}{100} \cdot x_1 > \frac{q}{100} \cdot \sum_{i=3}^{N} x_i$.

Hierzu folgendes Beispiel 5:
Es gilt die (17,6;50)-Regel → p=17,6% und q=50%

- Der Wert des Tabellenfeldes beträgt X=1000 €
- Der Wert des größten Einzelbeitrages sei x_1=650 €
- Der Wert des zweitgrößten Einzelbeitrages sei x_2=230 €
- Der Wert des drittgrößten Einzelbeitrages sei x_3=120 €

Da der Angreifer mit seinem Vorwissen unter der gewählten Annahme jeden Einzelbeitrag auf $\pm 50\%$ genau schätzen kann, wird er also x_3 mit einem Mindestwert von $\hat{x}_3 = 0{,}5 \cdot 120€ = 60€$ als untere Grenze annehmen. Damit kann der Angreifer mit dem zweitgrößten Einzelbeitrag unter Bezugnahme seines Vorwissens den größten Einzelwert wie folgt schätzen:

$\hat{x}_1 = X - x_2 - \hat{x}_3 = 1000€ - 230€ - 60€ = 710€$

$$\rightarrow \quad 100 \cdot \frac{\hat{x}_1 - x_1}{x_1} = 100 \cdot \frac{710€ - 650€}{650€} \approx 9{,}2$$

d.h. der Wert des größten Einzelbeitrags kann vom Angreifer auf $\approx 9{,}2\%$ genau geschätzt werden. Somit wäre dieses Tabellenfeld nach der (17,6;50)-Regel geheimzuhalten.
Die hergeleitete Bedingung:

$$\frac{p}{100} \cdot x_1 > \frac{q}{100} \cdot \sum_{i=3}^{N} x_i \quad \equiv \quad \frac{17{,}6}{100} \cdot 650€ > \frac{50}{100} \cdot \sum_{i=3}^{3} 120€$$ ist somit erfüllt.

Jedoch ist anzumerken, dass nur schwer beurteilt werden kann, wie genau ein Angreifer die restlichen Werte schätzen kann. Gegen Absprachen mehrerer Befragter die zum Gesamtzellwert beitragen, schützt diese Sicherungsmethode nicht. Es kann nur ein bestimmtes Niveau über das Vorwissen angenommen werden.

2.2. Die Sekundärsperrung

Da aggregierte Daten in Tabellen in einem additiven Zusammenhang stehen können (z.B. bei Tabellen mit Zwischen- und/oder Randsummen), müssen zusätzlich zu den primär gesperrten Tabellenfeldern noch sogenannte Sekundärsperrungen durchgeführt werden, um die Rückrechenbarkeit der sensiblen Daten durch Summen- oder Differenzbildung zu verhindern.

Folgendes Beispiel 6 soll dieses Problem verdeutlichen:

Dienstleister	Umsatz	Anzahl der Betriebe
davon: Finanzdienstleister	1000	12
Versicherungen	1200	4
Reinigungen	800	6
Banken	X	N_X
Unternehmensberatungen	Y	N_Y
Gesamt	T	N_T

Die Anzahl der Unternehmen je Dienstleistungssparte entsprechen den Fallzahlen aus Kapitel 2.1.1.

Die primäre Geheimhaltung werde nun nach der n=3 Fallzahlregel durchgeführt. Beispiel 6.1:
Die Fallzahl der Banken sei N_X=3
Die Fallzahl der Unternehmensberatungen sei N_Y=1.
Nach der Fallzahlregel n=1 darf der Umsatz der einzigen Unternehmensberatung nicht veröffentlicht werden, da dieser Einzelwert der Unternehmensberatung direkt zugeordnet werden kann. Das Tabellenfeld X ist somit ein sensibler Wert und damit primär zu sperren. Um die Rückrechenbarkeit von Y durch folgende Differenzbildung zu verhindern: T-1000-1200-800-X =Y, muss ein weiteres Tabellenfeld sekundär gesperrt werden, z.B. der Wert X.

2.2.1. Mindestumfang einer Sekundärsperrung

Auch Summen bzw. Differenzen geheimgehaltener Zellen (z.B. die Summe zweier primär gesperrter Zellen) lassen sich durch Differenzbildung ermitteln.
Beispiel 6.2: die Gesamtfallzahl N_T=22, N_X und N_Y seien jeweils gleich 1, dann hat die Tabellenspalte zwei Primärsperrungsfelder. Die Werte für X und Y dürfen nicht veröffentlicht werden.
Durch Differenzbildung kann nun die Größe des Summanden X+Y wie folgt berechnet werden: X+Y=T-1000-1200-800. Die einzige Bank X könnte nun mit ihrem Vorwissen über den eigenen Umsatz das primär gesperrte Feld Y durch einfache Differenzbildung ermitteln. Die Geheimhaltung der sensiblen Daten wäre mithin nicht gesichert. Deshalb ist eine zusätzliche Sekundärsperrung erforderlich. Die Summe gesperrter Zellen muss als mitveröffentlicht angesehen werden.
Beispiel 6.3: Es gelte die (1;85)-Dominanzregel
Der Wert des Gesamtumsatzes im Dienstleistungsgewerbe betrage T=4490, die Fallzahlen N_X und N_Y betragen jeweils drei. Die Einzelbeiträge zum Gesamtzellwert X belaufen sich auf: x_1=1300, x_2=70, x_3=30. Die Beitragsfolge zur Zelle Y sei y_1=35, y_2=30, y_3=25.
Dann wäre X aufgrund der (1;85)-Dominanzregel geheimzuhalten:

$$x_1 = 1300 > \frac{85}{100} \cdot \left(\sum_{i=1}^{3} x_i \right) = \frac{85}{100} \cdot (1300 + 70 + 30) = 1190 .$$

Der Umsatz der Unternehmensberatungen müsste nicht geheimgehalten werden, denn:

$$y_1 = 35 < \frac{85}{100} \cdot \left(\sum_{i=1}^{3} y_i \right) = \frac{85}{100} \cdot (35 + 30 + 25) = 76{,}5 .$$

Als Sekundärsperrungsfeld, um eine näherungsweise Aufdeckung des Umsatzes der größten Bank zu verhindern, kommt der Umsatz Y nicht in Frage, da er nicht groß genug ist.
Auch in der Summe der zwei gesperrten Zellen X+Y=Z würde der Umsatz der größten Bank dominieren.
Folgende Einzelbeitragsfolge würde entstehen:

z_1=1300 > z_2=70 > z_3=35 > z_4=z_5=30 > z_6=25

$$\rightarrow\ z_1 = 1300 > \frac{85}{100} \cdot \left(\sum_{i=1}^{6} z_i \right) = \frac{85}{100} \cdot (1300 + 70 + 35 + 30 + 30 + 25) = 1266{,}5 .$$

Folgende Sekundärsperrungsmöglichkeiten wären also denkbar: entweder statt des Unternehmensberatungsumsatzes ein anderes Tabellenfeld unterdrücken, oder den Unternehmungsberatungsumsatz zusammen mit einer weiteren Zelle sperren. Üblicherweise sollte die erste Möglichkeit bevorzugt werden, da hierbei der Informationsverlust geringer ist.

2.2.2. Regeln und Definitionen

Die Regeln für die primäre Geheimhaltung beschränken das Risiko der (näherungsweisen) Offenlegung von Einzelangaben. Sie dienen dazu, solche Zellwerte zu finden, bei denen dieses Risiko so hoch ist, dass die Einzelbeiträge auf eine nicht mehr akzeptable Weise abgeschätzt werden können.
Zum Beispiel wird nach der (1,85)-Dominanzregel ein Tabellenfeld geheimgehalten, wenn der Zellwert, der eine obere Grenze für den größten Einzelbeitrag darstellt, diesen um weniger als 17,6% übersteigt.

Definition[9]**:** Sei X ein Tabellenfeld mit N nach der Größe geordneten Einzelbeiträgen $x_1 \geq x_2 \geq ... \geq x_N$.

(a) Eine Linearkombination $S(X) = \sum_{i=1}^{N} w_i \cdot x_i$; w_i reell mit $w_1 \geq w_2 \geq ... \geq w_N$, heißt oberes lineares Sensitivmaß, wenn es natürliche Zahlen m, p_1, p_2 mit $0 < p_1 \leq p_2 \leq m$ gibt, so dass gilt: $w_i>0$; für $i<p_1$,
$w_i<0$; für $i>p_2$,
$w_i=w_m$; für alle $i \geq m$.

(b) Die monoton fallende Folge $w_1 \geq w_2 \geq ...$ heißt 'Folge der Gewichte von S(X)'.

(c) Ein Tabellenfeld X heißt sensitiv, wenn S(X)>0.

(d) Bei einem sensitiven Tabellenfeld bezeichnet man S(X) als Sensitivität des Tabellenfelds.

Gewichtfolgen lassen sich normieren, indem man den Betrag des m-ten Gliedes w_m durch das Sensitivmaß dividiert, so dass ab dem m-ten Glied gilt: $|w_i|=1$. Man spricht dann von einem normierten Sensitivmaß.
Ein Tabellenfeld X ist nach einer bestimmten Primärsperrungsregel genau dann geheimzuhalten, wenn es sensitiv ist, d.h. wenn gilt S(X)>0.

(1,k)-Dominanzregel
Nach der (1,k)-Dominanzregel ist ein Tabellenfeld geheimzuhalten, wenn:

$$x_1 > \frac{k}{100} \cdot X \quad \Leftrightarrow \quad x_1 - \frac{k}{100} \cdot X > 0$$

Durch Äquivalenzumformung mit $w_1 = 1 - \frac{k}{100}$ und $w_2 = w_3 ... = w_N = -\frac{k}{100}$

[9] Geurts, J.; 1992

ergibt sich aus Formel (A)[10]:

$$\left(1-\frac{k}{100}\right)\cdot x_1+\sum_{i=1}^{N}\left(\frac{-k}{100}\right)\cdot x_i>0 \quad \Leftrightarrow \quad S_{D(1,k)}(X):=\sum_{i=1}^{N}w_i\cdot x_i>0$$

ein Sensitivmaß $S_{D(1,k)}$ mit einer monoton fallenden Gewichtfolge, die ein positives und ansonsten negative

Glieder enthält: $$w_i=\begin{cases}1-k/100;\mapsto i=1\\ -k/100;\mapsto sonst\end{cases}$$

Allgemein gilt, dass für alle (n,k)-Dominanzregeln die Sensitivmaße mit

Gewichtfolgen: $$w_i=\begin{cases}1-k/100;\mapsto i=1,...,n\\ -k/100;\mapsto i=n+1,...,N\end{cases}$$ entsprechen.

(p,q)-Regeln

Die in Kapitel 2.1.2.5. beschriebenen (p,q)-Regeln entsprechen Sensitivmaßen mit Gewichtfolgen: $w_1=p/100$, $w_2=0$, $w_i=-q$ für $i\geq 3$.
Die oberen linearen Sensitivmaße mit der monoton fallenden Gewichtfolge $w_1\geq w_2\geq...$ besitzen eine besondere mathematische Eigenschaft, die sogenannte Subadditivität: Die Sensitivität einer Kombination zweier Tabellenfelder X und Y ist kleiner gleich der Summe ihrer Einzelsensitivitäten $S(X+Y)\leq S(X)+S(Y)$. Daraus folgt, dass wenn zwei Tabellenfelder nicht geheimgehalten werden müssen, auch die Kombination dieser beiden Zellwerte nicht als sensibel angesehen werden muss.
Beispiel 7: Es gelten die Annahmen von Beispiel 6.3. Die Beitragsfolge zum Tabellenfeld X (Umsatz der Banken) beträgt $x_1=1300$, $x_2=70$, $x_3=30$ und die Beitragsfolge zum Tabellenfeld Y (Umsatz der Unternehmensberatungen) $y_1=35$, $y_2=30$, $y_3=25$.

[10] Cox, L.H.; 1981

Aufgrund der (1,85)-Dominanzregel wäre X geheimzuhalten, da die normierte

Sensitivität[11]: $$S^*(X) = \frac{S_{D(1,k)}(X)}{|w_2|} > 0$$

$$\Updownarrow$$

$$S^*(X) = \frac{1}{w_2} \cdot \sum_{i=1}^{3} w_i \cdot x_i = \frac{100}{85} \cdot \left[\left(1 - \frac{85}{100} \right) \cdot x_1 - \frac{85}{100} \cdot x_2 - \frac{85}{100} \cdot x_3 \right] \approx 129{,}4 > 0 .$$

Weil $Y=y_1+y_2+y_3=90<S^*(X)$ kann Y nicht als Sekundärsperrung für X verwendet werden, wie auch schon im Beispiel 6.3 gezeigt wurde.
Für das kombinierte Tabellenfeld Z=X+Y mit der Beitragsfolge:

$z_1=x_1 > z_2=x_2 > z_3=y_1 > z_4=x_3 = z_5=y_2 > z_6=y_3$

gilt : $$z_1 = 1300 > \frac{85}{100} \cdot \left(\sum_{i=1}^{6} z_i \right) = \frac{85}{100} \cdot 1490 = 1266{,}5 .$$

Nach der (1,85)-Dominanzregel wäre demzufolge auch das kombinierte Tabellenfeld (X+Y) geheimzuhalten, weil es immer noch durch den größten Einzelbeitrag z_1 dominiert wird. Das aggregierte Tabellenfeld Z übersteigt den größten Einzelbeitrag (z_1) nur um 12,75%. Der größte Einzelbeitrag z_1 dürfte maximal 263 betragen.

2.2.3. Sekundärsperrungszyklen

In zweidimensionalen Tabellen mit Randsummen könnte man vermuten, dass die Aufdeckung geheimzuhaltender Werte genau dann nicht möglich ist, wenn in jeder Zeile und in jeder Spalte entweder kein Tabellenfeld oder zwei und mehr Zellen unterdrückt werden. Das diese Vermutung nicht immer zu einem Set von gesicherten Tabellenfeldern führt, soll folgendes Beispiel verdeutlichen:

[11] Jewett, R.; 1993

	1	2	3	4	Gesamt
1	X_{11}	X_{12}	X_{13}	2	13
2	X_{21}	2	X_{23}	7	13
3	3	X_{32}	8	X_{34}	14
4	4	X_{42}	3	X_{44}	18
Gesamt	11	10	19	18	58

An diesem Beispiel kann gezeigt werden, dass obwohl in jeder Zeile und jeder Spalte mindestens zwei Tabellenfelder gesperrt worden sind, sich der Wert der Zelle X_{12} exakt ermitteln lässt. Folgende Beziehungsgleichungen können aufgestellt werden:

$$X_{11} + X_{21} + X_{13} + X_{23} = 11 - 4 - 3 + 19 - 3 - 8 = 12$$
$$X_{11} + X_{12} + X_{13} + X_{21} + X_{23} = 13 - 2 + 13 - 2 - 7 = 15$$

Bildet man nun die Differenz zwischen den beiden Gleichungen, erhält man:

$$X_{12} = 3 .$$

Es lässt sich zeigen, dass eine gesperrtes Tabellenfeld sich genau dann nicht exakt ermitteln lässt, wenn es sich in einem bestimmten Zyklus von gesperrten Zellen befindet[12]. Als Zyklus bezeichnet man eine Folge von Null verschiedener Zellen der Form:

$$\{(i_0, j_0), (i_1, j_0), (i_1, j_1), (i_2, j_1), \ldots, (i_n, j_n), (i_0, j_n)\}$$

wobei alle i_k und j_l für k=0,1,...,n bzw. l=0,1,...,n (n: Länge des Zyklus) verschieden sind, d.h. $i_{k_1} \neq i_{k_2}$ wenn nicht $k_1 = k_2$ und $j_{l_1} \neq j_{l_2}$ wenn nicht $l_1 = l_2$.

Nachfolgendes Beispiel zeigt eine Tabelle, in der sich die gesperrten Zellen in einem Zyklus befinden[13]:

[12] Geurts, J. (1992, S. 10 ff.)
[13] Geurts, J. (1992, Tabelle 5, S. 10)

	1	2	3	4	Gesamt
1	X_{11}	X_{12}	2	2	13
2	2	X_{22}	X_{23}	7	13
3	3	2	8	1	14
4	X_{41}	2	X_{43}	3	18
Gesamt	11	10	19	18	58

Ein Spezialfall eines Zyklus ist ein Zellverbund, bei dem die gesperrten Zellen die Eckpunkte eines Quaders bilden. Auf das sogenannte Quaderverfahren wird in Kapitel 4 noch näher eingegangen. Dort werden auch n-dimensionale Tabellen behandelt.

2.2.4. Bestimmung von Wertebereichen

Aufgrund der additiven Zusammenhänge in einer Tabelle mit Randsummen lässt sich für die gesperrten Werte ein lineares Gleichungssystem aufstellen, was am folgenden Beispiel verdeutlicht werden soll[14]:

	1	2	Gesamt
1	X_{11}	X_{12}	7
2	X_{21}	X_{22}	3
3	3	3	6
Gesamt	9	7	16

$X_{11}+X_{12}=7$

$X_{21}+X_{22}=3$

$X_{11}+X_{21}=6$

$X_{12}+X_{22}=4$

mit $X_{ij} \geq 0$ für alle i und j

[14] Geurts (1992, Tabelle 10, S. 20)

bzw. in Matrix-Schreibweise:

$$\text{C X=b mit } C = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}, \quad X = \begin{bmatrix} X_{11} \\ X_{12} \\ X_{21} \\ X_{22} \end{bmatrix}, \quad b = \begin{bmatrix} 7 \\ 3 \\ 6 \\ 4 \end{bmatrix}$$

Um die Minimal- und Maximalwerte bestimmen zu können, wendet man ein Verfahren zur linearen Optimierung an, die sogenannte Simplex-Methode.
Nach dem Durchlauf dieses Verfahrens, würde man z.B. für X_{11} einen Minimalwert von X_{11}=3 und einen Maximalwert von X_{11}=6 erhalten.
Allgemein gilt, dass sich bei Tabellen mit positiven Zellwerten durch Verfahren der linearen Optimierung zu jeder gesperrten Zelle (i,j) ein oberer $\left(X_{ij}^{\max}\right)$ und unterer Grenzwert $\left(X_{ij}^{\min}\right)$ für den Zellwert bestimmen lässt. Im obigen Beispiel ist $X_{11}^{\min} = 3$ und $X_{11}^{\max} = 6$. Durch die Grenzwerte $\left(X_{ij}^{\max}\right)$ und $\left(X_{ij}^{\min}\right)$ für jede gesperrte Zelle (i,j) gegebene Intervall wird als Wertebereich der gesperrten Zelle (i,j) bezeichnet. Eine exakte Offenlegung gesperrter Tabellenfelder ist möglich, wenn der obere und untere Grenzwert für die gesperrte Zelle übereinstimmen.
Da Computerprogramme, die in der Lage sind Grenzwerte zu berechnen, frei verfügbar sind, ist die Ermittlung eines Intervalls für den jeweils gesperrten Wert nur mit geringem Aufwand verbunden. Sekundärsperrungen müssen daher so gewählt werden, dass die sich daraus ergebenden Grenzwerte für die primär geheimgehaltenen Werte es nicht ermöglichen, Verhältnisse der einzelnen zum sensiblen Wert beitragenden Erhebungseinheiten offenzulegen. Dies wäre beispielsweise möglich, wenn die Grenzwerte nicht einen gewissen Mindestabstand zum geheimgehaltenen Wert garantieren. Das Intervall zwischen kleinstem zulässigen oberen und größtem zulässigen unteren Grenzwert wird als Schutzintervall bezeichnet.

2.2.5. Informationsverlust

Durch die Zellsperrung entsteht für den Nutzer solcher Tabellen ein Informationsverlust. Die verschiedenen Möglichkeiten von Sekundärsperrungsmustern sollten so gewählt werden, dass der Informationsverlust möglichst gering gehalten wird. In Abhängigkeit von der Wahl des Maßes für den Informationsverlust werden Sperrmuster als günstig oder weniger günstig bewertet. Die folgende Tabelle zeigt, die für verschiedene Maße des Informationsverlustes resultierenden günstigen Sperrmuster.

Maß für den Informationsverlust ist...	Günstig sind Sperrmuster mit...
der Zellwert	kleiner gesperrter Wertsumme, auch wenn viele kleine Zellen gesperrt werden
die Fallzahl	geringer gesperrter Fallzahl, auch wenn viele kleine Zellen gesperrt werden
der Logarithmus des Zellwerts	Kompromiss zwischen minimaler Anzahl gesperrter Felder und minimaler gesperrter Wertsumme
für alle Zellen konstant, d.h. bei Auswahl der Sekundärsperrungen wird nicht nach dem Zellwert differenziert	geringer Anzahl gesperrter Felder, auch wenn deren Werte groß sind, oder es sich um (Zwischen-)Summen handelt

2.2.6. Das Sekundärsperrungsproblem als Optimierungsaufgabe

Der Widerspruch zwischen einem möglichst großen Schutzintervall der gesperrten Zellen und einem minimalen Informationsverlust führt zur Formulierung einer diskreten Optimierungsaufgabe.
Dabei soll der Informationsverlust, welcher durch die gesperrten Zellen entsteht, minimiert werden. Für eine zweidimensionale Tabelle mit m Zeilen, n Spalten und je einer Summenzeile und Summenspalte soll gelten:

$\sum_{i=1}^{m+1}\sum_{j=1}^{n+1} w_{ij} \cdot I_{ij} = Minimum$, wobei w_{ij} den festzusetzenden Informationsverlust

durch das Sperren der Zelle (i,j) und l_{ij} eine Indikatorvariable bezeichnet, die den Wert 1 annimmt bei zu sperrender Zelle (i,j) und sonst 0 ist. Die Gewichte w_{ij} werden auch als „Kosten“ der Zelle (i,j) bezeichnet. Die Minimierung der Zielfunktion wird unter folgenden Bedingungen durchgeführt:

(1) die Wertebereiche für alle Primärsperrungen sind groß genug
(2) die Wertebereiche aller gesperrten Zellen genügen bestimmten Kriterien
(z.B. positive Tabellen enthalten nur positive Werte)

Für jede gesperrte Zelle einer Tabelle mit Rand- und Zwischensummen kann mittels eines linearen Gleichungssystems ein Wertebereich ermittelt werden. Ein Sperrmuster, bei dem dieser Wertebereich nicht leer ist, erfüllt die Bedingung (2), d.h. das die Zelle auch besetzt sein muss, um einen Wertebereich bestimmen zu können. Gehören einem solchen Muster alle Primärsperrungen an und enthalten die Wertebereiche solcher Zellen nur Werte die außerhalb des definierten Schutzintervalls liegen, dann ist auch Bedingung (1) erfüllt und das Sperrmuster ist zulässig. Unter allen zulässigen Sperrmustern ist das auszuwählen, bei dem die Zielfunktion den kleinsten Wert annimmt, d.h. bei dem der Informationsverlust am geringsten ist.

2.2.7. Tabellen mit vorgegebenen Schätzintervallen

Es ist zu bedenken, dass die Tabellennutzer selbst zum Kreis der Berichtenden und zum Kreis der zu Schützenden gehören. Sie verfügen im Gegensatz zu den externen Nutzern über einen gewissen Grad an Vorwissen und haben berechtigtes Interesse, dass ihre Anonymität bei der Veröffentlichung nicht verloren geht. Die bisherigen Schutzmaßnahmen sind jedoch höchst unbefriedigend. Die Nutzer solcher Tabellen wissen aufgrund ihrer Erfahrung und ihrem Vorwissen viel mehr über diese Tabellen als nur, dass sie keine negativen Werte enthalten[15]. Es ist davon auszugehen, dass die geheimen Tabellenwerte auf bis zu plus minus 50% bekannt sind. Unterstellt man einen solch hohen Informationsgrad der Nutzer, dann reicht der hergeleitete Intervallschutz nicht mehr aus.

[15] Cox, L.H.; 1981

Folgendes Beispiel soll dieses Problem verdeutlichen:

100	80	180
90	1	91
190	81	271

Werden alle vier inneren Tabellenwerte zum Schutz des primär geheimzuhaltenden Wertes 100 gesperrt, so wird dieser durch ein Schutzintervall von range=80+1=81 bzw. einer relativen Spannweite von 81% gesichert. Bringt nun ein Nutzer dieser Tabelle sein Vorwissen in Form von Schätzintervallen ein, deren Intervallgrenzen um plus minus 50% vom tatsächlichen Wert abweichen, so liegt ihm folgende Tabelle vor:

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

Mit Hilfe dieser Tabelle und den Summenbeziehungen findet man dann für den primär geheimen Wert das Schutzintervall $99,5 \leq X_1 \leq 100,5$ oder eine relative Spannweite von 1%. Damit wäre der primär geheime Wert nahezu exakt bestimmt, obwohl lediglich die Randsummen veröffentlicht wurden. Zur Begründung des 1% Intervalls kann man aus der Tabelle mit den eingetragenen Sperrpositionen X_1, X_2, X_3 und X_4

X_1	X_2	180
X_3	X_4	91
190	81	271

die unbekannten inneren Tabellenwerte mit Hilfe der Randsummenwerte eliminieren und erhält eine Tabelle, mit einer einparametrigen Lösungsgesamtheit für die gesperrten Zellen. Der primär geheime Wert X_1 ist der zu schätzende Parameter.

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 vergleichen:

$50 \leq X_1 \leq 150 \rightarrow 50 \leq X_1 \leq 150$
$40 \leq 180 - X_1 \leq 120 \rightarrow 60 \leq X_1 \leq 140$
$45 \leq 190 - X_1 \leq 135 \rightarrow 55 \leq X_1 \leq 145$
$0{,}5 \leq 91 - (190 - X_1) \leq 1{,}5 \rightarrow 99{,}5 \leq X_1 \leq 100{,}5$

Der Nutzer wählt diejenigen Intervallgrenzen von X_1 aus, die X_1 am genausten festlegen, nämlich das Letzte der vier Intervalle. Nur mit der Information, dass eine nichtnegative Tabelle vorliegt, hätte der Nutzer folgende Intervalle bilden können:

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

Daraus hätte man das kleinste Intervall $99 \leq X_1 \leq 180$ gefunden, mit der relativen Spannweite 81%.
An diesem Beispiel kann man sehen, dass ein Vorwissen über die tatsächlichen Tabellenwerte von plus minus 50% genügt, um den zu schützenden Wert nahezu aufzudecken.

2.2.8. Heuristische Lösungsansätze

Da die Sekundärsperrung ein sehr komplexes Problem darstellt, arbeiten derzeit einige der programmierten Algorithmen heuristisch. Hierbei wird die Sekundärsperrung nicht idealerweise simultan für alle Primärsperrungen durchgeführt, sondern für die einzelnen Primärsperrungen nacheinander.
Zwischen folgenden drei Ansätzen wird derzeit unterschieden[16]:
Heuristik 1 – Eine n-dimensionale Tabelle mit Zwischen- und Randsummen wird in n-dimensionale Untertabellen zerlegt, so dass diese frei von Zwischensummen sind. Die Sekundärsperrung wird nacheinander in der Untertabellen-

[16] Gießing, S.; 1999

hierarchie absteigend durchgeführt. Innerhalb jeder Untertabelle wird die Sekundärsperrung für jede einzelne Primärsperrung durchgeführt. Zu jeder Primärsperrung wird ein n-dimensionaler Sperrquader ausgewählt, bei dem der primär gesperrte Wert einen der Eckpunkte bildet. Es kommen nur solche Sperrquader in Betracht, bei denen der Wertebereich der Primärsperrung hinreichend groß ist. Unter den möglichen Sperrquadern, die einen hinreichend großen Wertebereich der Primärsperrungsfelder garantieren, wird der Quader mit dem geringsten Informationsverlust ausgewählt.

Heuristik 2 – Der Reihe nach werden zu jeder Primärsperrung die dazugehörigen Sekundärsperrungen ausgewählt. Zunächst wird der sensible Wert so geändert, dass er außerhalb des Schutzintervalls liegt. Zum Ausgleich müssen auch andere Zellen ihre Werte ändern, um die Additivität der Tabelle zu erhalten. Die Lösung solcher Probleme wird unter Anwendung von Methoden der linearen Algebra gesucht und im Hinblick auf die Zielfunktion optimiert. Dieser Algorithmus neigt dazu, solche Lösungen zu suchen, bei denen die Abweichung vom Originalwert möglichst gering ist. Dadurch müssen relativ viele Zellen ihren Originalwert verändern, was nicht wünschenswert ist. Deshalb wird nach der Sekundärsperrung ein zweiter Durchlauf ausgeführt. Mittels veränderter Kosten w_{ij} in der Zielfunktion wird erreicht, dass nun große Zellen als Sperrpartner bevorzugt werden. Dabei sind nur die bereits im ersten Durchlauf gesperrten Zellen Sperrkandidaten. So wird geprüft, ob nicht auf einige kleinere Sperrpartner aus dem ersten Durchlauf verzichtet werden kann.

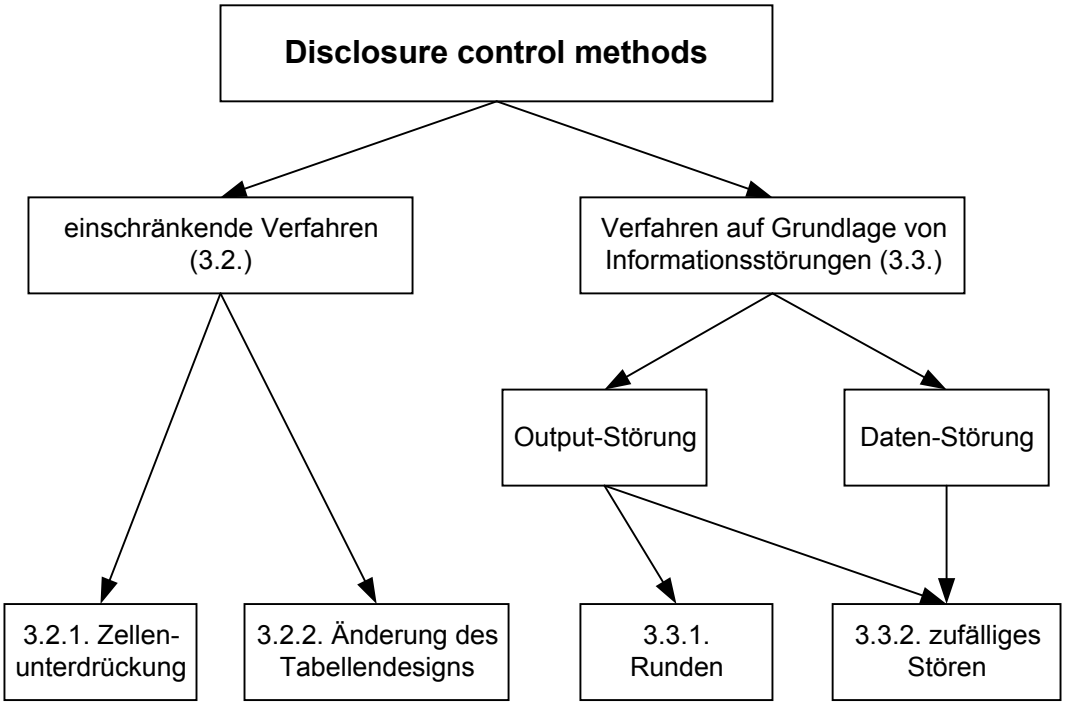
Heuristik 3 – Bei dieser Heuristik werden die Bedingungen (1) und (2) aus Abschnitt 2.2.6. für die Wertebereiche der gesperrten Zellen in Netzwerken umgesetzt. Als Lösung der Optimierungsaufgabe werden nur solche Sperrmuster bestimmt, bei denen die gesperrten Zellen einen Zyklus (vgl. Kapitel 2.2.3.) bilden. Allerdings lassen sich nur Tabellen mit höchsten zwei Dimensionen in ein Netzwerk umsetzen. Auch hier wird in einem zusätzlichen Durchlauf geprüft, ob eventuell auf kleine Sperrpartner aus dem ersten Durchlauf verzichtet werden kann.

Diese Heuristiken bilden die Grundlage für die maschinelle Durchführung des Zellsperrungsprozesses. Dabei findet Heuristik 1 bei dem Programm GHQUAR, Heuristik 2 bei CONFID und Heuristik 3 bei USBCSUP Anwendung.

3. Methoden zur Sicherung sensibler Daten

3.1. Einleitung

Es existieren zwei grundlegende Verfahren, um sensible Daten gegen Attacken potentieller Angreifer zu schützen. Bei der einen Möglichkeit wird der Informationsgehalt einer Tabelle reduziert. In der englischsprachigen Literatur werden diese Verfahren auch als 'restriction based disclosure control methods'(3.2.) bezeichnet. Der zweite Weg, Daten gegen ihre Aufdeckung zu schützen, wird als 'perturbation based disclosure control methods'(3.3.) bezeichnet. Folgende Übersicht ordnet die zur Anwendung kommenden Verfahren systematisch[17]:



Zur Auswertung der Verfahren und um einen Vergleich dieser zu ermöglichen, werden verschiedene charakteristische Merkmale herangezogen. Jedes einzelne Verfahren besitzt sowohl Vor- als auch Nachteile. In einigen Fällen kann es daher nötig sein, einzelne Techniken miteinander zu verknüpfen. Der Statistiker muss die für sein Problem beste Sicherungsmethode bestimmen.

[17] EUROSTAT 1996

Folgende Merkmalskriterien werden zur Beurteilung herangezogen:

1. Sicherheit: Eine wirkungsvolle Sicherungsmethode schützt die sensiblen Daten sowohl gegen exakte, als auch gegen näherungsweise Offenlegung. Das Sicherheitsniveau ist entsprechend hoch.
2. Robustheit: Eine Methode wird als robust bezeichnet, wenn es einem Angreifer trotz zusätzlichen Kenntnissen (z.B. in Form von Vorwissen) über die veröffentlichten Daten nicht gelingt, die sensiblen Werte zu enthüllen.
3. Flexibilität: Die Flexibilität einer Technik sollte so hoch, wie nur möglich sein. Ein Verfahren sollte sowohl mit qualitativen, wie auch mit quantitativen Attributen umgehen können.
4. Info.-gehalt: Eine hohe Aussagekraft (Minimierung des Informationsverlustes) ist eine zwingende Anforderung an geschützte Tabellen. Es soll nur soviel Information zurückgehalten werden, wie unbedingt nötig ist, um den Anforderungen des Datenschutzes zu genügen.
5. Kosten: Die Kosten sollten natürlich so niedrig wie möglich gehalten werden. Zwei Arten von Kosten entstehen hierbei. Erstens die Kosten, die zum Schutz der Daten aufgebracht werden müssen und zweitens die Kosten auf Seiten der Nutzer, um die Methoden zu verstehen und effektiv mit den Daten arbeiten zu können. Das Auftreten externer Effekte korreliert mit den Anforderungen an den Datenschutz. Volkswirtschaftlich muss für die Datensicherheit also ein hoher Preis gezahlt werden.

3.2. Methoden, die Tabellenfelder einschränken

3.2.1. Zellunterdrückung

Die Unterdrückung sensibler Tabellenfelder ist die am meisten verbreitetste Methode in der statistischen Geheimhaltung von Makrodaten. Die Zellenunterdrückungsmethode besteht aus der primären und der sekundären Sperrung. Um die Rückrechenbarkeit des primär gesperrten Wertes zu verhindern,

müssen noch zusätzliche Tabellenfelder sekundär unterdrückt werden[18]. Besitzt die Tabelle allerdings keine Rand- und/oder Zwischensummen, ist dieser Schritt der Sekundärsperrung nicht nötig, da der primär gesperrte Wert nicht durch Summen- oder Differenzbildung rückrechenbar ist.
Der primär gesperrte Wert, welcher vor der Öffentlichkeit aus Datenschutzgründen geheim gehalten werden muss, ist in der folgenden Tabelle mit einem x gekennzeichnet[19].

x	3	7
2	1	3
3	3	6
9	7	16

Tabelle 1

Der Wert der primär gesperrten Zelle x beträgt 4. Dies ist leicht durch Differenzbildung (z.B. 7-3=4) zu ermitteln.
Um die Tabelle ausreichend gegen Rückrechenbarkeit zu schützen, müssen nun noch zusätzliche Werte gesperrt werden (sekundär). In diesem vorgegebenen Szenario sind zwei Möglichkeiten denkbar.

x	x	7
x	x	3
3	3	6
9	7	16

Tabelle 2a

x	x	7
2	1	3
x	x	6
9	7	16

Tabelle 2b

Um beurteilen zu können, welcher dieser beiden Möglichkeiten den Vorzug zu geben ist, sind Regeln notwendig. Es gibt eine Vielzahl solcher Regeln. Folgende zwei Möglichkeiten werden in der Literatur am häufigsten verwendet:

- Minimierung der totalen Anzahl der geschützten Zellen, und/oder
- Minimierung der totalen Werte, die hinter den geschützten Zellen verborgen sind.

In diesem Beispiel schützen beide Tabellen (2a und 2b), die selbe Anzahl an Zellen, nämlich 4. Jedoch ist der Wert unterschiedlich. Bei Tabelle 2a beträgt er

[18] Sande, G. (S. 33-41) 1984
[19] Robertson, D. 1991

10 und bei 2b 13. Deshalb ist nach der zweiten Regel die Möglichkeit a, der von b vorzuziehen. Oftmals wird gesagt, dass die Minimierung der totalen Anzahl der gesperrten Zellen wichtiger ist, als die Minimierung der totalen Zellenwerte[20]. Zu wissen, dass 95% des Umsatzes eines Unternehmens mit nur einem Produkt erreicht werden, ist aufschlussreicher, als zu wissen, dass 5% des Umsatzes durch das restliche Sortiment erreicht wird. Die Minimierung der totalen Werte ist wichtiger bei kleinen gesperrten Tabelleneinträgen.
Die Ermittlung von Wertebereichen bei dieser Methode wurde ausführlich in Kapitel 2.2.4. behandelt.

Auswertung dieses Verfahrens:

1. Sicherheit		hoch
	(a) exakte Aufdeckung	nein
	(b) teilweise Aufdeckung	ja, durch Wertebereiche
2. Robustheit		niedrig
3. Flexibilität		hoch
4. Aussagekraft		mittelmäßig
5. Kosten		mittelmäßig

3.2.2. Veränderung des Tabellendesigns

Es existieren zwei Möglichkeiten, das Design einer Tabelle zu verändern und somit sensible Daten zu unterdrücken[21].

- Reduzierung der Tabellengröße (auch als Rolling-up bekannt)
- Recoding der Tabellencharakteristik

3.2.2.1. Reduzierung der Tabellengröße

Hierbei ist die verbreitetste Variante aus einer Tabelle mit n Zeilen und m Spalten eine weniger detailierte Tabelle mit nur (n-1) Zeilen und/oder (m-1) Spalten zu erzeugen. Die einfache Idee hinter dieser Technik besteht darin, dass bei

[20] Kelly, J.P./ Golden, B.L./ Assad, A.A. (S.397-417) 1992
[21] Dalenius, T. (S. 20-5,6) 1988

einer Reduzierung der Tabelleneinträge sich auch der veröffentlichte Informationsgehalt verringert. Folgendes Beispiel soll dies veranschaulichen[22]:

	W	X	Y	Z	Total
A	12	10	**2**	6	30
B	24	**3**	5	8	40
Total	36	13	7	14	70

Tabelle 3

	W	X+Y	Z	Total
A	14	12	6	30
B	24	8	8	40
Total	36	20	14	70

Tabelle 4

Beim Addieren von Spalte X mit Spalte Y aus Tabelle 3 entstand die kleinere (2x3) Tabelle 4. Kritisch ist bei dieser Methode zu beachten, dass es nicht immer sinnvoll sein wird Spalten zu addieren, die unter logischen Gesichtspunkten nur unabhängig voneinander eine sinnvolle Information liefern. Die gesperrten Tabellenfelder sind zwar auf jeden Fall gesichert, jedoch werden die aggregierten Daten unbrauchbar.

3.2.2.2. Recoding der Tabellencharakteristik

Diese Methode beschreibt die Möglichkeit die Attribute einer Ausgangstabelle in Kategorien zusammenzufassen. Die Kategorien, welche in der neuen Tabelle entstehen, müssen so gewählt werden, dass der sensible Wert innerhalb einer größeren Gruppe von Attributen versteckt wird[23]. Es sollte jedoch darauf geachtet werden, dass die Kategorien nicht zu groß gewählt werden, da sonst der Informationsverlust zu hoch wäre.

[22] Eurostat (S. 24-26) 1996
[23] Willenborg, L./ deWaal, T. (S. 176) 2001

Hierzu folgendes Beispiel:

Alter	<12	12	13	14	15	16	17	18	19	<19	Total
nicht kriminell	10	7	6	5	2	4	3	3	5	12	57
kriminell	**3**	2	2	1	1	2	1	1	3	5	21
Total	13	9	8	6	3	6	4	4	8	17	78

Tabelle 5

Der sensible Wert (3) ist hervorgehoben. Nach dem Recoding ist dieser Wert nicht mehr rückrechenbar, wie Tabelle 6 zeigt:

Alter	unter 14	14-17	über 17	Total
nicht kriminell	23	14	20	57
kriminell	**7**	5	9	21
Total	30	19	29	78

Tabelle 6

Tabelle 6 zeigt, dass der sensible Wert der unter 12-jährig Kriminellen in der zusammengefassten Gruppe der unter 14-Jährigen verschwindet. Eine exakte Aufdeckung dieses Wertes ist nun nicht mehr möglich. Diese Tabelle 6 zeigt allerdings auch, dass der Informationsverlust sehr hoch ist. Damit haben diese Verfahren des Tabellenredesigns erhebliche Nachteile gegenüber dem Verfahren der Zellenunterdrückung, da hier sämtliche Einträge von der Unterdrückung der wenigen sensiblen Daten betroffen sind und verändert werden. Durch das Zusammenfassen von Zeilen und/oder Spalten, kann die gesamte Aussagekraft einer Statistik verloren gehen. Vergleiche mit z.B. anderen Ländern oder Regionen werden dadurch fast unmöglich.

Auswertung des Verfahrens:

1. Sicherheit		hoch
	(a) exakte Aufdeckung	nein
	(b) teilweise Aufdeckung	ja, abhängig von den Änderungen
2. Robustheit		moderat
3. Flexibilität		hoch
4. Informationsgehalt		moderat bis niedrig
5. Kosten		niedrig

3.3. Verfahren auf der Grundlage von Informationsstörungen

3.3.1. Runden

Beim Runden werden kleine Werte gegen Aufdeckung geschützt. Die Idee dahinter ist, jeden Wert N_{ij} in einer zweidimensionalen Tabelle zum nächsten ganzzahligen vielfachen Wert auf- oder abzurunden.
Ist $N_{ij} > h \cdot b$, dann kann N_{ij} geschrieben werden als:

$N_{ij} = h \cdot b + r_{ij}$ [24]

r_{ij} ist der Restwert mit $0<r_{ij}<b$. Der Rest r_{ij} wird nun aufgerundet zu b (ganzzahlige Basis) oder abgerundet zu 0. Ein Beispiel soll diese Definition illustrieren: der zu rundende Zellenwert N_{ij} sei 83 und b=5, dann ist h (das größte ganzzahlige Vielfache) gleich 16. Daraus folgt: $N_{ij} = 83 = 16 \cdot 5 + 3$.

Der Rest r_{ij} ist somit gleich 3 und ist entweder aufzurunden zu b=5 (Resultat N_{ij}=85) oder abzurunden auf 0, dann wäre N_{ij}=80. Dies ist abhängig von der verwendeten Rundungsmethode. In diesem Kapitel werden drei Methoden des Rundens vorgestellt. Die Unterschiede dieser Methoden soll folgendes Beispiel aufzeigen.

1	4	0	15	20
15	10	10	20	55
2	10	10	3	25
2	6	15	12	35
20	30	35	50	135

Tabelle 7a

Zunächst muss für jede Rundungsmethode die Basis b festgelegt werden. Je größer b gewählt wird, desto sicherer ist der Schutz und desto geringer ist der Informationsgehalt. Eine größer gewählte Basis bedeutet, dass der definierte Bereich vom Rest r_{ij} ($0<r_{ij}<b$) sich vergrößert und somit die Differenz zwischen dem gerundeten Wert N_{ij}' und dem sensiblen Wert N_{ij} ansteigt.
Der Wert für den Informationsverlust L kann kalkuliert werden:

$L = \sum_i \sum_j N_{ij} - N'_{ij}$.

[24] Fellegi, I.P. (S. 123-133) 1975

3.3.1.1. Konventionelles Runden

Bei diesem Beispiel wird eine Basis b=5 benutzt. N_{ij} wird gerundet zu 0 oder 5. Endet n_{ij} auf 1 oder 2 wird auf Null abgerundet und bei 3 oder 4 zu 5 aufgerundet. Werte die schon auf 0 oder 5 enden, bleiben unverändert ($N_{ij}=N_{ij}'$). Aus Tabelle 7a wird somit die folgende Tabelle 7b[25]:

0	5	0	15	20
15	10	10	20	55
0	10	10	5	25
0	5	15	10	35
20	30	35	50	135

Tabelle 7b

Das konventionelle Runden ist ein sehr einfaches Verfahren. Hierbei ist die Abweichung zwischen dem realen und dem gerundeten Wert sehr gering, was mit einem geringen Informationsverlust einhergeht.
In diesem Beispiel beträgt der Informationsverlust:

$$L = \sum_i \sum_j \left| N_{ij} - N'_{ij} \right| = 1+1+2+2+2+1+2 = 11.$$

Im weiteren wird sich zeigen, dass der Informationsverlust beim wahllosen und beim kontrollierten Runden größer ist, weil bei diesen Methoden nicht immer mathematisch korrekt gerundet wird.
Wie man an Tabelle 7b sehen kann, hat das konventionelle Runden auch seine Grenzen. Die Randsummen stimmen nicht mehr mit den gerundeten Werten N_{ij}' überein. Eine korrekte Summenbildung ist nicht mehr möglich, ohne Veränderung der Ausgangsdaten. Das zweite Problem ist, dass die Nullen, welche durch Runden entstehen, nicht sicher sind. Folgendes simple Beispiel zeigt dies deutlich:

0	0	4
0	0	4
4	4	8

Tabelle 8

[25] Eurostat; 1996

Da in diesem Beispiel die Randsummen nicht stimmen, ist es offensichtlich, dass es sich hierbei um keine „echten“ Nullen, sondern um gerundete Nullen handelt. Jede Null war vor dem Runden eine 2. Andere Möglichkeiten (wie z.B. 1 und 3) kommen nicht in Betracht, da eine 3 auf 5 aufgerundet werden würde.

3.3.1.2. Wahlloses Runden

Hierbei wird eine fixe Basis b bestimmt. N_{ij} wird mit einer Wahrscheinlichkeit p abgerundet und mit einer Wahrscheinlichkeit 1-p aufgerundet. Die Wahl der Wahrscheinlichkeit macht das zufällige Runden vorurteilslos. Das Runden wird nach folgendem Schema vorgenommen[26]:

	die letzte Ziffer von N_{ij} ist:					
	0	1	2	3	4	5
dann runde zu '0' mit p=	1	4/5	3/5	2/5	1/5	0
und zu '5' mit 1-p	0	1/5	2/5	3/5	4/5	1

Tabelle 9

Diese Tabelle sagt aus, dass wenn z.B. ein Wert N_{ij} auf die Ziffer 1 endet, dieser mit einer Wahrscheinlichkeit von 4/5 zu Null abgerundet und mit 1/5 zu 5 aufgerundet wird. $E(N_{ij}')= 4/5\ (N_{ij} – 1) + 1/5\ (N_{ij}+4) = Nij$, wenn die letzte Ziffer 1 ist.
Bei der vorgegebenen Tabelle kann beispielsweise folgendes entstehen:

0	**0**	0	15	20
15	10	10	20	55
0	10	10	5	25
0	**10**	15	**15**	35
20	30	35	50	135

Tabelle 10

Die drei hervorgehobenen Werte sind entgegen ihrer Wahrscheinlichkeit rein zufällig entgegengesetzt gerundet worden. Durch dieses zufällige mathematisch falsche Runden ist diese Tabelle besser geschützt, als die beim konventionellen Runden. Ein potentieller Angreifer kann sich hierbei nicht sicher sein, ob auch

[26] Nargundkar, M.S./ Saveland, W. (1972)

immer mathematisch richtig gerundet wurde. Somit ist eine exakte Aufdeckung gesperrter Werte, im Gegensatz zum konventionellen Runden, nicht mehr möglich. Der Nachteil besteht jedoch darin, dass auch hier die Randsummen nicht mehr mit den Tabelleneinträgen überein stimmen. Hinzu kommt, dass der Informationsverlust viel höher, als beim konventionellen Runden ist. In diesem konkreten Beispiel beträgt er: $L = \sum_i \sum_j \left| N_{ij} - N'_{ij} \right| = 1+4+2+2+2+4+3 = 18$.

3.3.1.3. Kontrolliertes Runden[27]

Das kontrollierte Runden wurde entwickelt, um die Nachteile vom konventionellen und vom zufälligen Runden zu beseitigen und gleichzeitig die Vorteile zu nutzen. Beim kontrollierten Runden ist es möglich, den Wert der Zelle N_{ij} beliebig zu einem der beiden möglichen ganzzahligen Vielfachen von b zu runden[28]. Rundet man die Zellwerte vorausschauend entweder ab oder auf, so kann man sicherstellen, dass die additiven Zusammenhänge der Tabelle erhalten bleiben. Dabei kann zusätzlich der Informationsverlust L minimiert werden. Um beim Runden die Variante auszuwählen, welche den Optimalzustand sicherstellt, kommt ein spezielles Linearprogramm zum Einsatz. Dieses Programm wird in der englischsprachigen Literatur als Transportation Model bezeichnet. Hierbei werden die unterschiedlichen Rundungsalternativen miteinander verglichen und das Optimum ausgewählt.
Aus dem Ausgangsbeispiel entsteht folgende Tabelle[29]:

0	5	0	15	20
15	10	10	20	55
0	10	10	5	25
5	5	15	10	35
20	30	35	50	135

Tabelle 11

Der hervorgehobene Wert 5, stellt den einzigen Unterschied zum konventionellen Runden dar. In dieser gesicherten Tabelle können korrekte Zeilen- und

[27] Greenberg, B. (1990)
[28] Fischetti, M./ Salazar, J.J. (1996)
[29] Cox, L.H. (1987)

Spaltensummen gebildet werden. Der Informationsverlust beträgt: $L = \sum_i \sum_j \left| N_{ij} - N'_{ij} \right| = 1+1+2+2+3+1+2 = 12$.

Der Unterschied zum minimalen Informationsverlust beim konventionellen Runden ist gering.

Bei sehr großen Tabellen kann jedoch ein Problem mit der Rechenzeit auftreten. Es muss ständig überprüft werden, ob die additiven Zusammenhänge noch korrekt bestehen. Möglicherweise müssen schon gerundete Werte rückwirkend noch einmal neu gerundet werden.

3.3.1.4. Vergleich der Rundungsmethoden

		konventionelles Runden	zufälliges Runden	kontrolliertes Runden
Sicherheit		mittel	hoch	hoch
	(a) exakte Aufdeckung	ja	nein	nein
	(b) näherungsweise Aufdeckung	ja, abhängig von der Basis	ja, abhängig von der Basis	ja, abhängig von der Basis
Robustheit		mittel	mittel	mittel
Flexibilität		mittel	mittel	mittel
Info-gehalt		mittel	niedrig	mittel
Kosten		niedrig	niedrig	hoch

3.3.2. Zufälliges Stören[30]

Bei dieser Methode wird eine zufällige Störvariable e zum wahren Zellwert X_{ij} hinzuaddiert. Anschließend wird der wahre Wert durch den Zufallswert X'_{ij} ersetzt. $X'_{ij}=X_{ij}+e$, mit E(e)=0 und var(e)=b[31]

Für die zufällige Störvariable e ist eine Verteilung zu bestimmen. Anschließend ist ein Set von Störvariablen e und die dazugehörige Wahrscheinlichkeit auszuwählen. Daraus entsteht folgender Erwartungswert E(e):

$E(e)_{=e} \sum [p(e) \cdot e] = 0$ und eine bekannte Varianz von var(e)=b. Je größer die

[30] Evans/ Zayatz/ Slanta (1996)
[31] Adam/ Wortmann (1989)

Varianz gewählt wird, desto besser sind die sensiblen Daten vor einer Offenlegung geschützt und desto größer ist die Verzerrung der Tabellenwerte.
Der Betrag der Störung, gleichbedeutend mit dem Informationsverlust, ist die Differenz zwischen dem originalen und dem gestörten (verfälschten) Wert: $L = \sum_i \left| \sum_j X_{ij} - X'_{ij} \right|$.

Bei der praktischen Anwendung werden für die Störvariablen e Werte definiert, die sich symmetrisch um die Null bewegen (-k bis +k). Die Störvariable kann für jede Zelle die gleiche sein, aber auch von Zelle zu Zelle variieren.
Folgendes Beispiel verdeutlicht dieses Verfahren[32]:

e	-2	-1	0	1	2
p(e)	1/9	2/9	3/9	2/9	1/9

Tabelle 12

1	3	4	8
5	2	2	9
7	9	8	24
13	14	14	41

Tabelle 13A: X_{ij}

1-2	3-1	4^+/-0
5-1	2^+/-0	2+1
7^+/-0	9+1	8+2

Tabelle 13B: X_{ij}+e

0	2	4	8
4	2	3	9
7	10	10	24
13	14	14	41

Tabelle 13C: X'_{ij}

[32] Eurostat; 1996

Man kann entweder die Störvariable zu jeder Zelle (inkl. der Randsummen) hinzu addieren, oder nur zu den „inneren“ Tabellenfeldern (ohne die Randsummen; wie in diesem Beispiel). In beiden Fällen sind die additiven Zusammenhänge der Tabelle zerstört.
In Tabelle 12 sind die Störvariablen e und die jeweils dazugehörigen Eintrittswahrscheinlichkeiten p(e) angegeben. Tabelle 13A zeigt die Originalwerte. 13B zeigt das zufällige Aufaddieren der Störvariablen unter Berücksichtigung ihrer auftretenden Häufigkeiten. Tabelle 13C zeigt eine Möglichkeit eines sicheren Tableaus. Ein spezieller Fall tritt in Zelle (1,1) auf. Nach mathematisch korrekter Addition der Störvariable, müsste der Wert $x'_{11} = -1$ betragen. Da negative Zellwerte nicht definiert sind, werden solche Tabellenfelder immer gleich Null gesetzt.

Auswertung des Verfahrens:

1. Sicherheit		hoch
	(a) exakte Aufdeckung	nein
	(b) teilweise Aufdeckung	ja, abhängig von der Störvariable
2. Robustheit		moderat
3. Flexibilität		moderat
4. Informationsgehalt		moderat
5. Kosten		niedrig

4. Quaderverfahren[33]

4.1. Einführung

Im folgenden wird das vom Landestatistikamt Nordrhein-Westfalen[34] entwickelte Quaderverfahren zur Wahrung der Geheimhaltung aggregierter Daten beschrieben. Das Verfahren sichert sensible Daten in Tabellen mit Zwischen- und Randsummen (auch unterteilte Tabelle) gegen zu genaue Rückrechnung. Es bietet einen Intervallschutz, d.h., dass ein Angreifer einen gesperrten Tabellenwert nur außerhalb des definierten Intervalls schätzen kann. Das Quaderverfahren wurde insbesondere zum Schutz von sehr umfangreichen Tabellen entwickelt. Bei höherdimensionalen Tabellen erweist sich schon die Prüfung, ob ein geheimer Tabellenwert bereits gesichert ist oder nicht, als sehr zeitaufwendig. Zur Lösung ist ein lineares Gleichungssystem mit den geheimen Werten als Unbekannte erforderlich. Zur Vereinfachung dieses Problems bietet sich eine Reduktion auf unabhängige Einzelgleichungen mit dem Differenzverfahren an. Es wird hierbei für jede Dimension geprüft, ob die geheime Zelle die Einzige ist, die zu einer Summe beiträgt oder nicht. Es wird also untersucht, ob sich der geheime Wert durch Differenzbildung mit einem Summenwert und den anderen zu dieser Summe beitragenden Werte berechnen lässt oder nicht. Nach diesem Prüfungsverfahren werden nur solche geheimen Werte als gesichert angesehen, die einem n-dimensionalen Quader angehören.

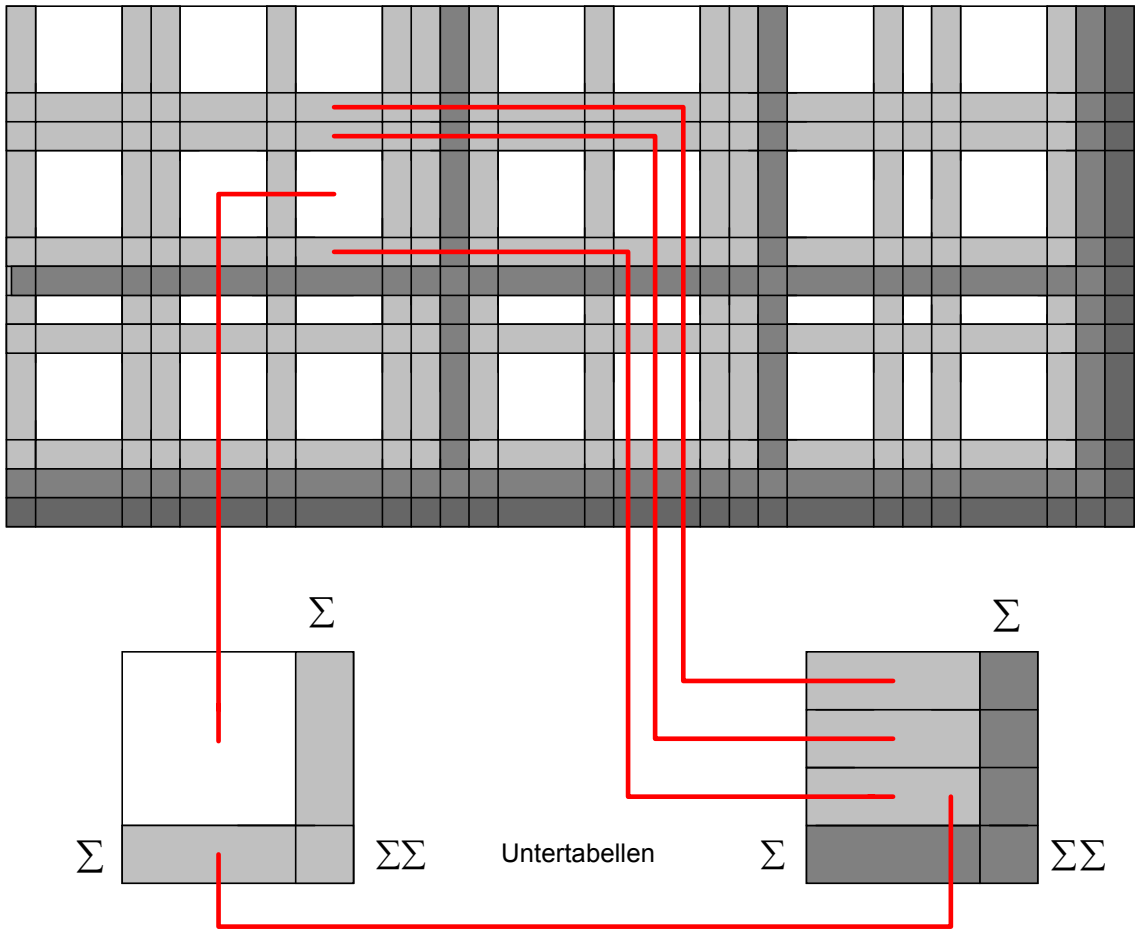
4.2. Grundlegende Probleme der sekundären Geheimhaltung

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, gestaltet sich relativ einfach. In Kapitel 3 wurden diese Verfahren bereits vorgestellt. Die Kriterien, nach denen eine Sekundärsperrung durchgeführt wird, finden beim Quaderverfahren nach folgenden Prioritäten ihre Anwendung. In erster Linie wird die Anzahl der Sekundärsperrungen minimiert und erst in zweiter Linie wird eine

[33] Repsilber, D.; 1999
[34] Statistische Studien und Analysen NRW Ausgabe 3/2000

kleine Wertsumme der Sekundärsperrungen angestrebt. Die Durchführung der sekundären Geheimhaltung gestaltet sich bei durch Zwischensummen untergliederten Tabellen sehr viel komplizierter. Diese Teilgesamtheiten einer Gesamttabelle werden im folgenden als Untertabelle bezeichnet. Durch die Behandlung von einzelnen Untertabellen wird das Problem der Geheimhaltung der Gesamttabelle in eine Vielzahl kleiner überschaubarer Teilprobleme zerlegt. Wie bei der Geheimhaltung einer Einzelangabe (Fallzahl n=1 vgl. Kapitel 2), die durch die Einbindung in eine Tabelle gefährdet ist, verhält es sich mit ganzen Untertabellen. Diese sind in die Aggregationsstufenhierarchie einer Gesamttabelle eingeordnet. Wie sich solche Untertabellen gegenseitig beeinflussen können, zeigt folgende Darstellung[35]:

Gesamttabelle einer zu sichernden Statistik



Wechselbeziehungen zwischen Untertabellen

[35] Darstellung siehe Repsilber, D.; 1999

Folgendes Beispiel soll dieses Problem deutlich machen:

Heraussperren					
Kreise	**Gruppe**				
	A	B	C	D	Gesamt
1		8 240			8 240
2	3 240	**2**[1.)] **187**	**33**[2.)] **184**	67 1782	105 2393
3		3 16			3 16
4		87 448			87 448
Bezirk	3 240	**100**[3.)] **1191**	**33**[3.)] **184**	67 1782	203 3397

obere Zeile: Anzahl; untere Zeile: Betrag

1.) geheimzuhaltender Wert

2.) Unterdrückung zur Vermeidung der Rückrechenbarkeit des geheimen Wertes

3.) leere Tabellenfelder erzwingen Summensperrungen als zusätzliche Sperrpartner

Bei schwach besetzten Tabellen, wie in diesem Beispiel, kann es sein, dass sich mit den inneren Tabellenfeldern (d.h. gleiche Aggregationsstufe) kein Karree zur Sicherung eines geheimen Wertes finden lässt. Hier muss man auf die Randsummen ausweichen. Dadurch entstehen neue geheime Werte im inneren einer Tabelle der nächsthöheren Aggregationsstufe, die dann in dieser Tabelle gesichert werden müssen. Dadurch erhöht sich in jeder Aggregationsstufe die Anzahl der notwendigen Sperrungen um ein Vielfaches. Das primär gesperrte Feld (2,B) der Beispieltabelle lässt sich zwar durch das Sperren von (2,C) gegen Rückrechnung mittels Differenzbildung in der Zeile 2 schützen, bezüglich der Spalten fehlen jedoch besetzte sperrbare Tabellenfelder. Zumindest wenn man davon ausgeht, dass „Nullsperrungen“ nicht zulässig sind. Die Sekundärsperrung (2,C) kann nur durch die Summensperrung (Bezirk,C) gesichert werden. Daraus ergibt sich das Karree (2,B), (2,C), (Bezirk,B), (Bezirk,C) mit zwei erzwungenen Randsummensperrungen. Im Laufe des Sicherungsvorgangs der Gesamttabelle können Sekundärsperrungen in Tabellen höherer Aggregationen auftreten. Diese findet man in den zugehörigen Tabellen

niedrigerer Verdichtung als Summensperrungen wieder. Hier sind unter Umständen zusätzliche Sperrungen im Inneren der Tabelle nötig, wie folgende Tabelle beispielhaft zeigt:

Hineinsperren					
Kreise	**Gruppe**				
	A	B	C	D	Gesamt
1	11 7760	8 240	4 57	117 4154	140 12211
2	**3**[4.)] **240**	**2**[1.)] **187**	**33**[4.)] **184**	**67**[2.)] **1782**	105 2393
3	322 1723	**3**[2.)] **16**	18 115	**8**[1.)] **258**	351 2412
4	116 842	87 448	21 439	4 86	228 1815
Bezirk	**452**[3.)] **10565**	100 1191	**76**[3.)] **795**	196 6280	824 18831

obere Zeile: Anzahl; untere Zeile: Betrag

1.) geheimzuhaltender Wert

2.) Unterdrückung zur Vermeidung der Rückrechenbarkeit des geheimen Wertes

3.) Löschung höherer Hierarchiestufen

4.) Löschung in höherer Hierarchiestufe erzwingt Sperrung

Die Sicherung jeder Untertabelle für sich allein betrachtet, stellt eine aus Sicht der Gesamttabelle unzulässige Idealisierung dar. Der Grund dafür liegt darin, dass Sperrungen in Untertabellen höherer Aggregationsstufen immer auch Sperrungen in den zugehörigen Untertabellen niedrigerer Verdichtung bedeuten. Nur wenn aufgrund günstiger Tabellenfeldbelegung alle Sperrungen, primär wie sekundär, sich in jeder Gliederung auf das unterste Niveau beschränken, können die Untertabellen unabhängig voneinander betrachtet werden. In allen anderen Fällen sind die Untertabellen voneinander abhängig. Aus diesem Grund bietet sich ein zweistufiges heuristisches Verfahren an. In der ersten Stufe wird mittels des Quaderverfahrens die Geheimhaltung in jeder einzelnen Untertabelle gesichert. Die zweite Stufe umfasst den gegenseitigen Abgleich aller Untertabellen. Da jede Sekundärsperrung in einer Untertabelle höherer Verdichtung, immer auch eine Summensperrung in einer der zugehörigen Untertabellen niedrigerer Aggregationsstufen bedeutet, wird jeweils mit der Bearbeitung der

Untertabellen höchster Aggregationsstufen begonnen. Absteigend werden alle Aggregationsstufen abgearbeitet, bis alle Untertabellen gesichert sind. Dabei werden die fortlaufend in die Gesamttabelle eingetragenen Sekundärsperrungen der anderen Untertabellen mitberücksichtigt (Untertabellenabgleich). Es sei bereits an dieser Stelle erwähnt, dass auch noch eine andere Möglichkeit besteht, eine durch Zwischensummen unterteilte Tabelle so zu organisieren, dass sie mit dem Quaderverfahren bearbeitet werden kann. Diese Möglichkeit wird erst im Kapitel 4.6.2., welches sich mit überlappenden Tabellen befasst, eingehender diskutiert.

4.3. Vermeidung eindeutiger Rückrechenbarkeit

4.3.1. Einführung des Quaderkonzepts

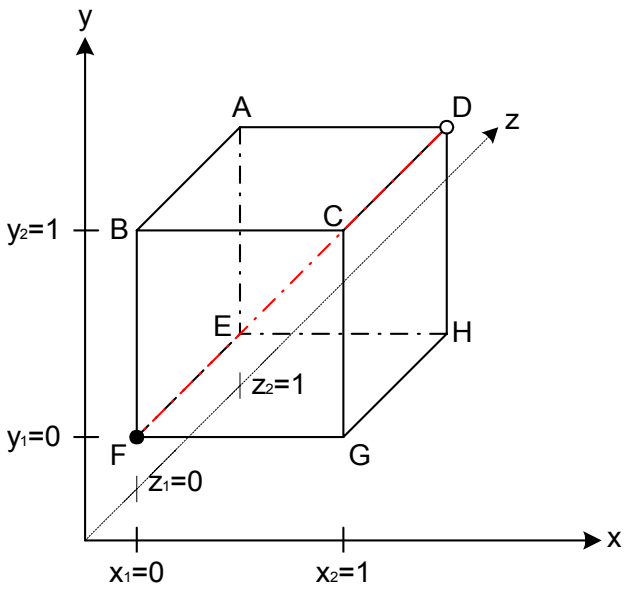
Das Quaderverfahren bietet einen hinreichenden Schutz gegen eindeutige Rückrechenbarkeit, wenn die Prüf- und Sperrfunktion in einem Schritt vereinigt sind. Es werden die primär geheimzuhaltenden Werte einer n-dimensionalen Untertabelle überprüft, ob sie einem n-dimensionalen Quader mit lauter gesperrten Werten angehören (Prüffunktion) und sichert sie gegebenenfalls durch Sperren noch offener Quaderwerte (Sperrfunktion).

Das Quaderverfahren ist besonders von Bedeutung:

- weil es bei nicht zu großen Tabellen sowohl maschinell als auch manuell durchgeführt werden kann; es besteht direkte manuelle Überprüfbarkeit
- weil es n-dimensionale Tabellen von der Größenordnung 1000000 Tabellenfelder mit geringem Rechenzeitaufwand gegen Rückrechnung sichern kann
- weil es ein optimales Verfahren ist, das für nur einen zu sichernden Wert die kleinste Anzahl von Sekundärsperrungen auswählt.

4.3.2. Allgemeine Regeln und Definitionen

In dem folgenden dreidimensionalen Quader ist der geheime Tabellenwert (F) durch weitere geheime Werte (Primär- und/oder Sekundärsperrungen) in den Ecken eines Quaders gesichert worden.



	x	y	z
F	0	0	0
E	0	0	1
B	0	1	0
A	0	1	1
G	1	0	0
H	1	0	1
C	1	1	0
D	1	1	1

Um den geheimen Wert F zuerst in seiner Ebene y_1 zu schützen, wird das Karree $K(y_1)=\{(x_1,y_1,z_1), (x_2,y_1,z_1), (x_2,y_1,z_2), (x_1,y_1,z_2)\}$ ausgewählt. Da eine Rückrechnung auch über die dritte Dimension erfolgen kann, wird noch ein weiteres Karree $K(y_2)$ als Projektion von $K(y_1)$ aufgesucht und alle noch offenen Werte gesperrt.

Zur Übertragung dieses Konzepts auf n-dimensionale Tabellen bedarf es folgender zusammenfassender Definitionen:

1. Ein Tabellenwert heißt zu einem anderen diametral, wenn sich die Indizes beider Werte in jeder Dimension unterscheiden. Im vorangegangenen Beispiel wären das die Punkte (A-G), (B-H), (C-E), (D-F).
2. Die Gesamtheit aller n-dimensionaler indizierter Tabellenwerte, die durch jeweils zwei zueinander diametrale Werte festgelegt sind, heißen n-dimensionale Quader.
3. Ein durch n-Dimensionen indizierter geheimer Wert (Einzelangabe → Fallzahl=1 oder nicht) heißt quadergesichert, wenn er zur Gesamtheit eines n-dimensionalen Quaders mit lauter von Null verschiedenen gesperrten Werten gehört, die keine Einzelangaben sind (Fallzahl>1).

Um eine minimale Anzahl von Sekundärsperrungen zu erzielen, werden folgende Regelungen getroffen:

1. Von allen Quadern, die mit den primär zu sichernden Werten gebildet werden können, soll derjenige ausgewählt werden, welcher schon die meisten Primärsperrungen beinhaltet. Stehen dann noch mehrere Sperrquader zur Auswahl, ist derjenige zu bevorzugen, der die minimale Wertsumme an Sekundärsperrungen sicherstellt.
2. Besteht ein Sperrquader aus mehr als einer Einzelangabe, so muss noch ein zweiter Quader zum Schutze der anderen Einzelangaben erzeugt werden, welcher diese Zellen ausschließt.

4.3.3. Behandlung von Einzelangaben

Nach Definition 3 sollten Einzelangaben keine „Sicherungspartner“ in einem n-dimensionalen Quader sein, weil diese Einzelangaben eindeutig berechnet werden können, wenn nicht noch ein weiterer Sperrquader zur Sicherung ausgewählt wird.

Folgende (Unter-)Tabelle soll dieses Problem verdeutlichen:

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

obere Zeile: Fallzahl (n=1); untere Zeile: Zellwert

= geheimzuhaltender Wert.

Mit Ausnahme der Randsummen sind alle Werte nur einem Merkmalsträger zugeordnet. Deshalb ist allein durch den Quader (1A), (1B), (2A), (2B) der Wert der primär zu sperrenden Zelle nicht gesichert. Jeder Merkmalsträger der anderen Quaderzellen kann mit dem Vorwissen über seinen eigenen Beitrag den Wert des Tabellenfeldes 1A durch Differenzbildung mit dem Summenwert und den anderen hier zunächst als offen angenommenen nicht zum obigen Quader gehörigen Werte bestimmen. Durch Auswahl eines zweiten Quaders zum Schutze von 1A, der die Einzelzellen des ersten Quaders außer 1A nicht enthält, z.B. (1A), (1C), (3A), (3C), wird die eindeutige Rückrechenbarkeit von (1A) verhindert. Der Angreifer mit dem Vorwissen von Zelle 3C könnte zwar die anderen Werte seines Quaders 3A und 1C berechnen, jedoch fehlt ihm das Wissen über die Werte des anderen Quaders. Somit hat der Angreifer 3C keine Möglichkeit, den Wert der Zelle 1A zu ermitteln. Führt man diese Prozedur für jeden möglichen Angreifer durch, müssen alle Werte der Tabelle primär gesperrt werden. Diese Erkenntnis bestätigt die Aussagen in Kapitel 2.1.a, wonach Tabellenwerte mit der Fallzahl n=1 nicht gesichert sind.

4.4. Herleitung der Quader-Indexformel

Die Koordinaten eines n-dimensionalen geheimzuhaltenden Wertes F seien durch: $f=(f_1, f_2, f_3,..., f_n)$ gegeben. Ein dazu diametraler Tabellenwert D habe die Indizes $d=(d_1, d_2, d_3,..., d_n)$ mit $d_i \neq f_i$ für i=1, 2, 3,..., n

Das bedeutet, dass die zum selben Gliederungskriterium i gehörenden Indizes d_i und f_i voneinander verschieden sind. Der zu den beiden zueinander diametralen Werten D und F gehörige Quader ist die Gesamtheit aller Tabellenwerte Q, die durch $(q_1, q_2, q_3,..., q_n)$ indiziert sind, wobei gilt:

$Q_i \rightarrow$ entweder f_i oder d_i; i=1, 2, 3,..., n

Da jeder Indexwert q_i zum Gliederungskriterium i zwei Werte annehmen kann (f_i oder d_i), besteht der Quader aus 2^n Tabellenwerten. Um alle 2n Quaderwerte aufsuchen zu können, 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 lediglich auf die Indizes an. Der Normquader lässt sich beschreiben durch:

$(B_1(k), B_2(k),... ,B_i(k),..., B_n(k))$, k=0, 1, 2,..., 2^n-1 wobei mit $B_i(k) \rightarrow$ entweder 0 oder 1; i=1, 2, 3,..., n eine binäre Variable eingeführt wird und k die Nummer des betrachteten Normquaderwertes ist. Jeder Normquader ist durch ein n-Tupel von Nullen und Einsen indiziert, die als Binärstellen der Nummer des betreffenden Wertes aufgefasst und in eine natürliche Dezimalzahl k umkodiert werden können. Ist z.B. (0,1,0,0,1) das Index-5-Tupel eines Normquaderwertes einer fünfdimensionalen Tabelle, so ist die Nummer dieses Normalquaders:

$$k=01001_{bin} = 0 \cdot 2^4 + 1 \cdot 2^3 + 0 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0 = 9_{dez}.$$

Demnach haben die den Normquader fixierenden zueinander diametralen Werte (0,0,0,0,0) und (1,1,1,1,1) die Nummern k=0 und k=31.

Man erhält auf diese Weise eine ganz bestimmte mit Null beginnende Nummerierung aller Quaderwerte und zwar so, dass die zum Gliederungskriterium i gehörige Binärstelle des k-ten Quaderwertes 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 $\{f_i\}$ 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 Wert (f_1, f_2, f_3,…, f_n) identifiziert und den dazu diametralen Normquaderwert mit Nummer $k=2^n-1$ entsprechend (1,1,1, …, 1) mit dem 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 $q_i(k)=f_i$ immer $B_i(k)=0$ und $q_i(k)=d_i$ immer $B_i(k)=1$ zugeordnet ist:

Für i=1, 2, ..., n gilt: $Q_i(k)=f_i \leftrightarrow B_i(k)=0$ $q_i(k)=d_i \leftrightarrow B_i(k)=1$.

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

$$q_i(k) = f_i + B_i(k) \cdot (d_i - f_i)$$

für i=1,2,3,..., n und k=0,1,2,...,2^n-1 zu berechnen, wobei $B_i(k)$ die i-te Binärstelle des Laufindex-Wertes k zum i-ten Gliederungskriterium bezeichnet. Die Anwendbarkeit dieser Quader-Index-Formel setzt voraus, dass die Ausprägungen der Gliederungsmerkmale ganzzahlig sind.

4.5. Zum Intervallschutz

4.5.1. Bestimmung der Spannweite geheimer Tabellenwerte

Wie schon in Kapitel 2 beschrieben, kann ein Angreifer mittels eines linearen Gleichungssystems den wahren sensiblen Tabellenwert durch Verfahren der linearen Optimierung eingrenzen. Aus der Differenz zwischen dem maximalen und dem minimalen Wert für das sensible Tabellenfeld kann dann die Spannweite berechnet werden ($range_k$=max X_k -min X_k (k=1, 2, 3,..., r)). Diese

Spannweite kann nur Bruchteile von Prozent eines geheimen Wertes X_k betragen. Der notwendige Mindestschutz wäre dann nicht mehr gewährleistet.

4.5.2. Spannweitenabschätzung mittels eines Quaders

Mittels eines n-dimensionalen Quaders können Spannweiten geheimer Werte bestimmt werden, die höchstens so groß, wie die mit linearer Optimierung berechneten sind.

Definition: Ein Quaderwert X heiße gerade indiziert, wenn die Anzahl der Indizes (q_1, q_2, q_3,... q_n) die mit den entsprechenden Indizes des diametralen Wertes D übereinstimmen, gerade ist, anderenfalls heiße er ungerade indiziert. Das bedeutet, dass ein Quaderwert gerade indiziert ist, wenn die Summe der Binärstellenwerte seiner Quaderwertnummer k gerade ist (siehe Quader-Indexformel).

Hierzu folgendes Beispiel: Im Beispiel 4.3.2. sind die Werte F, C, H und A gerade indiziert. Die Normquaderindizes lauten (0,0,0), (1,1,0), (1,0,1) und (0,1,1,). Jeder dieser Werte besitzt eine gerade Anzahl von Einsen, die mit D (1,1,1,) übereinstimmen. Demzufolge sind die Werte B, D, E und G ungerade indiziert.

Um ein lineares Gleichungssystem für die 2^n Quaderwerte X als Unbekannte aufstellen zu können, hat man gemäß der Summenvorschrift der Untertabelle, für jedes Gliederungskriterium i über alle Indexausprägungen zu summieren. Weil jeder Quaderwertindex nur zwei Werte annehmen kann, tragen auch nur zwei Quaderwerte X, X' zur jeweiligen Randsumme bei. Daher haben alle linearen Gleichungen des Quaders folgende Gestalt: X+X'=∑. Diese Summe bezeichnet die Quaderwertsumme des i-ten Gliederungskriteriums. Das heißt, die Randsumme abzüglich aller anderen Summanden des i-ten Gliederungskriteriums, die nicht zum Quader gehören.

Für jeden Quaderwert X und für jedes der n-Gliederungskriterien lässt sich eine Gleichung aufstellen. Jede dieser Gleichungen enthält genau 2 der 2^n Quaderwerte. Es lassen sich demnach: $\frac{2^n \cdot n}{2} = n \cdot 2^{n-1}$ Gleichungen entwickeln. Bei

Tabellen, die drei und mehr Dimensionen besitzen, erhält man mehr Gleichungen als Unbekannte. Davon sind jedoch nur 2^n-1 voneinander unabhängig.

Für den dreidimensionalen Quader 4.3.2. ergibt sich beispielsweise durch Summenbildung über das erste Gliederungskriterium bei festen, zweiten und dritten Index z.B.: $E_{0,0,1}$+$H_{1,0,1}$+Summe aller nicht zum Quader gehörigen inneren Tabellenwerte mit festen, 2. und 3. Index (x,0,1) =(Randsumme,0,1).

Ist nun die Quadergleichung X gerade indiziert, so ist X' ungerade indiziert, weil sich beide Werte nur im Summationsindex unterscheiden. Das bedeutet, dass X' einen diametralen Indexwert d_i im Summationsindex i mehr oder weniger hat als X. Schätzt ein Angreifer den gerade indizierten Wert X durch die Gleichung: $\hat{X} = X + \varepsilon \geq 0$, so muss er den ungerade indizierten Wert X' auf $\hat{X}' = X' - \varepsilon \geq 0$ schätzen, damit die Annahmen der Quadergleichung erfüllt sind. Diese beiden Beziehungen gelten für alle Quaderwerte mit demselben ε-Wert als Schätzfehler. Beispielsweise kann ein gerade indizierter Wert Z eines Quaders von X ausgehend „erreicht" werden, indem ein Index von X nach dem anderen in den entsprechenden Index von Z umgesetzt wird. So erhält man aufeinanderfolgende Quaderwerte X, X', Y, Y',..., Z, Z', von denen je zwei benachbarte Werte immer durch eine Quadergleichung der Gestalt X+X'=Σ miteinander verknüpft sind: $X+X'=\Sigma_{XX'}$, $X'+Y=\Sigma_{X'Y}$, $Y+Y'=\Sigma_{YY'}$,..., $Z+Z'=\Sigma_{ZZ'}$. In dieser Folge von Quadergleichungen haben immer je zwei benachbarte Gleichungen einen Quaderwert als Summanden gemeinsam. Für einen Angreifer gilt nach jeder einzelnen Indexumbesetzung immer $\hat{X} = X + \varepsilon \geq 0$ bei gerader und $\hat{X}' = X' - \varepsilon \geq 0$ bei ungerader Indizierung.

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

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

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

$$\hat{Z} = Z + \varepsilon, \qquad \hat{Z}' = Z' - \varepsilon.$$

Obige Gleichungen gelten für alle Quaderwerte X, X' mit demselben Schätzfehler ε. Sie enthalten genau einen Parameter (ε). Da ε frei wählbar ist, bedeutet

dies einen hinreichenden Schutz gegen Rückrechnung geheimer Werte. In welchen Grenzen ε frei gewählt werden kann, hängt von den Quaderwerten ab.

Werden nichtnegative Tabellenwerte unterstellt, müssen auch die Schätzwerte der geheimen Quaderwerte nicht negativ sein. Somit können positive ε-Werte höchstens so groß, wie der kleinste ungerade indizierte Quaderwert minX' und negative ε-Werte betragsmäßig höchstens so groß, wie der kleinste gerade indizierte Quaderwert minX sein. Sind also negative Schätzwerte auszuschließen, so muss für $\varepsilon=\varepsilon_1\geq 0 \rightarrow \varepsilon_1 \leq \min X'$ und für $\varepsilon=\varepsilon_2<0 \rightarrow |\varepsilon| \leq \min X$ sein. Das heißt:

$$\hat{X} \in \left[X - \min X; X + \min X'\right], \quad \hat{X}' \in \left[X' - \min X'; X' + \min X\right]. \qquad \text{(A)}$$

Im Inneren einer Untertabelle sind gemäß obiger Ungleichungen zwei Fehlschranken zugeordnet, sein kleinster gerade indizierter Wert und sein kleinster ungerade indizierter Wert. Die Intervalllänge (A) ist daher für alle Quaderwerte, egal ob gerade oder ungerade indiziert, die gleiche: Range=minX'+minX.

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.

2 (g_1g_2)	2 (g_1d_2)
2 (d_1g_2)	2 (d_1d_2)

Der Sicherungsquader umfasst dem gemäß das Pivotelement im Inneren der Tabelle mit den Indizes (g_1; g_2) und den Aggregationsstufen (1;1), das wegen 0+0+1+1=2 gerade indiziert ist, die beiden Randsummenwerte mit den Indizes (g_1; d_2), (d_1; g_2) und den Aggregationsstufen (1;2), (2;1), die daher ebenfalls gerade indiziert sind (für das erste der beiden Felder gilt 0+1+1+2=4), und das Summenfeld (d_1; d_2) mit den Aggregationsstufen (2;2), das ebenfalls gerade indiziert ist (1+1+2+2=6). Der kleinste gerade indizierte Wert ist demnach der (einzige) Tabellenwert selbst. Da in dem hier betrachteten Quader offensichtlich kein ungerade indizierter Quaderwert existiert, gibt es auch keinen kleinsten dieser Werte, so dass das Intervall der gerade indizierten (und daher auch aller)

Schätzwerte in (A) 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 Quaderwertnummer k und seiner Aggregationsstufen gerade ist, anderenfalls ist er ungerade indiziert (Aggregationsstufen mit 1 beginnend aufsteigend durchnummeriert).

Laut dieser Definition behalten Schätzwertintervalle und die Spannweite der geheimen Quaderwerte, auch für Quader mit Randsummen, ihre Gültigkeit. Ist der range-Wert nicht größer als der, welcher mittels linearer Optimierung berechnet wurde, dann hat man ein Quaderauswahlkriterium das einen hinreichenden Intervallschutz bietet. Mit dem zu definierenden Prozentwert q werden nur solche Quader zur Sicherung eines geheimen Wertes X ausgewählt, für die: $\frac{100 \cdot range}{X} > q$ gilt. Wählt man beispielsweise den Prozentwert q=85%, so lässt die Auswahlregel nur solche Quader zur Sicherung eines primär geheimen Wertes X zu, deren Spannweite bezogen auf den zu sichernden Wert X größer als 85% ausfällt. Es werden also nur solche Quader ausgewählt, deren kleinster gerade indizierter und kleinster ungerade indizierter Wert in ihrer Summe größer als $0{,}85 \cdot X$ sind.

Anmerkungen

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 Schutzschranke 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.

4.6. Sicherung von Tabellen mit gemeinsamen Aggregaten[36]

4.6.1. Tabellenübergreifende Geheimhaltung

In der Praxis treten nicht nur einzelne voneinander unabhängige Tabellen auf, sondern auch mehrfach durch Zwischensummen unterteilte Tabellen, die sich einander überlappen. Hierbei existieren Aggregate die mehrere Tabellen gemeinsam haben. Bei der Sicherung würde hierbei kein neues Problem auftauchen, wenn es gelänge, die Überlappungsbereiche beim Sperren zu vermeiden. In der Praxis hat sich jedoch gezeigt, dass Sperrungen in Überlappungsbereichen nicht auszuschließen sind. Daraus ergibt sich die zwingende Notwendigkeit, dafür zu sorgen, dass mehreren Einzeltabellen gemeinsam angehörende Aggregate in allen Einzeltabellen den gleichen Geheimhaltungsstatus besitzen. Für die Sicherung solcher voneinander abhängiger Tabellen kommt daher nur eine gemeinsame Bearbeitung durch gegenseitigen Abgleich in Frage, analog zum Abgleich der Untertabellen. Alle Tabellen, die zu einem Set aneinander abzugleichender Einzeltabellen gehören, sollten i.d.R. gleichzeitig veröffentlicht werden. Eine später erstellte Veröffentlichungstabelle, die Überlappungen mit bereits veröffentlichten Tabellen besitzt, kann nur dann gesichert werden, wenn der Abgleich mit allen in Frage kommenden „Vorgängertabellen“ ausschließlich Sperrungen in der „Nachzüglertabelle“ hervorbringen, sonst nicht. Den Eingabebestand für dieses Abgleichsverfahren erhält man, indem man eine n-dimensionale Tabelle nach der anderen in den Datenbestand überträgt und mehrfach vorkommende Sätze löscht. Dabei werden die Gliederungsmerkmale

[36] de Wolf, P.P.; 1999

der Einzeltabellen als neue Tabellenmerkmale in den Bestand übernommen. Der Überlappungsbereich zeigt sich dadurch, dass dort mehr Gliederungsmerkmale auftreten als in jeder Einzeltabelle. Bei der Bearbeitung einander überlappender Tabellen können Übersperrungen auftreten, weil bei jedem neuen Durchlauf auch die Sekundärsperrungen überprüft und ggf. gesichert werden. Das führt bei Einzelangaben oft zu weiteren Sperrungen, weil zwar jeder primär geheime Wert durch die nur zu seinem Schutz gesetzten Sekundärsperrungen vollkommen gesichert ist, nicht aber unbedingt umgekehrt die Sekundärsperrungen durch Einzelangaben oder durch andere primäre Sperrungen.

4.6.2. Rückführung von überlappenden auf vollständige Tabellen

4.6.2.1. Rückrechenbarkeit sicherer Untertabellen

Wie festgestellt, wurde auch der gegenseitige Untertabellenabgleich in Bezug auf die Summensperrungen dadurch erreicht, dass die gesamte Untertabellenhierarchie in mehreren Schritten so lange durchlaufen wurde, bis keine weiteren Sekundärsperrungen mehr auftraten. Dieses Vorgehen ist zwar notwendig, nicht jedoch hinreichend, für die Sicherung der Gesamttabelle. Folgendes Gegenbeispiel soll dies verdeutlichen: Rückrechenbarkeit über mehrere, in sich sichere und einander abgeglichene Untertabellen.

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

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

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

Ursächlich für die Rückrechenbarkeit geheimer Werte ist die Aufteilung des durch die Summenvorschriften der Gesamttabelle gegebenen linearen Gleichungssystems auf die einzelnen Untertabellen. Bei Summensperrungen sind die Teilsysteme, deren Untertabellen gemeinsam zu denselben Randsummen beitragen, nicht unabhängig voneinander. Sie müssen somit bei der Sicherungsprozedur gemeinsam bearbeitet werden. Das Phänomen der möglichen Rückrechenbarkeit einander überlappender Tabellen (d.h. Tabellen die gemeinsame Tabellenfelder besitzen) ist keine Besonderheit des Quaderverfahrens. Dieses Problem tritt unabhängig vom Untertabellensperralgorithmus auf. Aus diesem Grund bieten gesicherte Untertabellen, welche mit den vorgestellten Methoden der linearen Optimierung gesichert wurden, keinen hinreichenden Schutz.

4.6.2.2. Aufstockung der Tabellendimension

Will man gekoppelte Untertabellen zusammenfassen, muss die Tabellendimension aufgestockt werden. Dies soll an folgender, mehrfach durch Zwischensummen untergliederter, eindimensionaler Tabelle verdeutlicht werden:

$a_1, a_2, a_3,...$	Σ_1	$b_1, b_2, b_3,...$	Σ_2	...	$v_1, v_2, v_3,...$	Σ_m	$\Sigma\Sigma$

In dieser Tabelle werden die Elemente der untersten Aggregationsstufe zu ihren Zwischensummen $\sum_i$ aufaddiert. Diese Zwischensummen werden dann ebenfalls zusammengefasst und ergeben die Gesamtsumme $\sum\sum$. Aufgrund des Assoziativgesetzes und der Kommutativität der Addition könnte man diese Zusammenfassung auch wie folgt vornehmen:

1. Es werden die ersten Elemente der untersten Aggreagtionsstufe zur Zwischensumme $\sum^*_1$ aufaddiert.
2. Nun geschieht das Gleiche mit den zweiten Elementen der untersten Aggregationsstufe zu $\sum^*_2$ usw.
3. Nun werden die Zwischensummen $\sum^*_j$ (mittlere Aggregationsstufe nach dieser neuen Gliederung) zur Gesamtsumme $\sum\sum$ (höchste Aggregationsstufe) aufaddiert.

Man erhält folgende zweidimensionale Tabelle, welche nicht mehr durch Zwischensummen untergliedert sind und als vollständig bezeichnet werden.

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

Bei der Transformation in eine solche Form, kann es vorkommen, dass die Anzahl der Kategorien bezüglich der aufzustockenden Gliederung in den einzelnen Untertabellen nicht übereinstimmen. Beispielsweise könnten 10 Elemente a_i zur Summe $\sum_1$ und nur 5 Elemente b_i zur Summe $\sum_2$ usw. beitragen. Ist das der Fall, so müssen die nicht zusammenpassenden Gliederungen durch leere Kategorien (Dummy-Kategorien) ergänzt werden. Die Umstrukturierung einer n-dimensionalen Tabelle geschieht ganz analog, 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 ist.

Definition: Eine Statistiktabelle heißt vollständig, wenn die Addition von Tabellenwerten über jedes Gliederungskriterium (über jeden Index) zu genau einer Summe, der Randsumme, führt.

Eine vollständige Tabelle ist somit eine aus der Untertabellenhierarchie herausgelöste für sich selbst betrachtete Tabelle. Die (Aggregat-)Dimension einer vollständigen Tabelle ergibt sich als Summe der höchsten Aggregationsstufen bezüglich jedes, durch die ursprüngliche Tabelle gegebenen Gliederungskriteriums vermindert, um die Anzahl dieser Gliederungskriterien. Beispielsweise entsteht aus einer zweidimensionalen Statistik (z.B. Wirtschaftssystematik mit 7 Aggregationsstufen und regionale Gliederung mit 4 Aggregationsstufen) eine vollständige neundimensionale Tabelle. Die Aufstockung der Dimension führt bei realen Statistiktabellen in der Regel zu sehr umfangreichen, hochdimensionalen, vollständigen Tabellen. Diese sind gegenüber den ursprünglichen, mehrfach durch Zwischensummen unterteilten Tabellen, durch Einfügen zusätzlicher Summen unterschiedlicher Aggregation und durch Eintragung strukturgebender Tabellenfelder, der Dummies, erweitert worden. Dabei kommt einigen Dummy-Werten dieselbe Bedeutung zu, wie den strukturellen Nullen. Sie können nicht zur Sicherung geheimer Werte gesperrt werden! Wenn diese Aufstockung von dem für die Wahrung der Geheimhaltung sensibler Daten verantwortlichen Fachstatistiker unter ausschließlicher Verwendung von Realdaten durchgeführt wird, oder wenn die zur Sicherung anstehenden Tabellendaten von vornherein, d.h. nach fachlichen Gesichtspunkten bereits so strukturiert werden, dass nur noch vollständige Tabellen vorliegen, dann kann das Quaderverfahren ohne weitere Vorbereitungen angewendet werden; es bietet dann einen hinreichenden Schutz gegen zu genaues Rückrechnen primär geheimer Werte.
Da eine Dimensionsaufstockung nur angezeigt ist, wenn Sperrungen in den Randsummen der Untertabellen auftreten, und da Sperrungen im Rand erfahrungsgemäß seltener vorkommen als im Inneren von Untertabellen, bietet sich ein zweistufiges Vorgehen an, wonach im ersten Schritt, alle primär geheimen Tabellenwerte auf unterstem Aggregationsniveau ohne Aufstockung der Dimension gesichert werden und wonach erst im zweiten Schritt die noch verbliebenen Sicherungen, die zu Randsperrungen führen, nach der Dimensionsaufstockung erfolgen.

Die Beispieltabelle 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. Folgende Abbildung zeigt, wie die Rückrechenbarkeit der Beispieltabelle behoben werden kann.

X_1 X_4	0 0	X_1 X_4	X_2 0 [1.]	0 X_5	X_2 X_5	X_3 0	0 X_6	X_3 X_6	10 [2.] 20 [2.]
20	0	20	X_2	X_5	X_2+X_5	X_3	X_6	X_3+X_6	30
0 0	0 0	0 0	X7 0	0 X_9	X_7 X_9	X_8 0	0 X_{10}	X_8 X_{10}	40 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

[1.] Wird als einziger Wert gesperrt, wenn Nullen sperrbar sind.

[2.] Werden keine Nullen akzeptiert, müssen die beiden Randsummenwerte 10 und 20 gesperrt werden.

Anmerkung: Wenn die Sperrung „[1.]“ 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.

4.7. Anwendungsmöglichkeiten und Schlussbemerkungen

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, sowie Art und Grad der Sicherung. Diese Faktoren können bei der Bearbeitung von Geheimhaltungsproblemen nicht vollständig unabhängig voneinander realisiert werden. Die gegenseitige Abhängigkeit der Faktoren schränkt die bestehenden Auswahlmöglichkeiten der Faktorkategorien also weitgehend ein.

Die nachfolgende Übersicht enthält die in Betracht kommenden Faktoren, mit ihren Kategorien und Bemerkungen mit anwendungsrelevanten Hinweisen.

Quaderverfahren für n-dimensionale Tabellen			
Faktor	Faktorkategorie		
Tabellentyp	Wertetabelle ohne Zwischensummen	Wertetabelle mit Zwischensummen	überlappende Tabellen
Bemerkung	ohne Zwischensummen gegeben oder Aufstockung	Untertabellenhierarchie	Randschranken einsetzen
Vorinformation	Tabelle enthält positive und neative Werte	Tabelle enthält nur positive Werte	es existieren Schätz-intervalle
Bemerkung	Nullwerte haben keine Sonderstellung	Nullen nur in einer Quaderteilgesamtheit oder Intervallschutz	zu kleine Schätzintervalle verhindern Geheimhaltung
Tabellen-organisation	Untertabellenhierarchie mit Abgleich	Vervollständigung durch Aufstockung	Zusammenführung überlappender Tabellen in gemeinsame Datenbestände
Bemerkung	Verfeinerung zu Tabellenteilen mit Rand- ohne Zwischensummen	Vergröberung der Gesamttabelle ohne Zwischensummen	alle Tabellen gemeinsame Aggregate werden nur einmal aufgeführt
Sicherungsart	Sekundärsperrungen gegen		Verfälschungen durch Zufallsfelder
	eindeutige Rückrechnung	zu genaue Rückrechnung	
Bemerkung	ohne Intervallschutz	mit Intervallschutz	gerade indiziert (+); ungerade indiziert (-)
Sicherungsgrad	nur notwendiger Schutz bei Tabellenabgleich	hinreichender Schutz nur bei Tabellen ohne Zwischensummen	
Bemerkung	Abgleich von überlappenden Tabellen	zu erreichen durch Vermeidung von Summensperrungen und durch Dimensionsaufstockung	

5. Vergleich der zur Verfügung stehenden Software[37]

5.1. Vorstellung der vorhandenen Programme

Derzeit befinden sich fünf Programme zur Durchführung der Sekundärsperrung weltweit im Einsatz:

Programm-name	Programm-entwicklung durch	Sekundärsperrungs-methode	Maß für den Informationsverlust
GHQUAR	LDS NRW	Quaderverfahren (vgl. Heuristik 1)	Zahl der gesperrten Felder und Wertsummen
USBCSUP	US Bureau of Census	Netzwerktechnik (vgl. Heuristik 3)	gesperrte Wertsumme
CONFID	Statistics Canada	Lineare Optimierung (vgl. Heuristik 2)	Logarithmus der gesperrten Werte
ACSSuprs	Sande and Associates Inc.	Lineare Optimierung (vgl. Heuristik 2)	Logarithmus der gesperrten Werte
Tau-ARGUS	CBS Niederlande	Diskrete lineare Optimierung	gesperrte Wertsumme

In den weiteren Vergleich wird das Programm ACSSuprs nicht einbezogen, da der zu Grunde liegende Sekundärsperrungsalgorithmus der selbe ist, wie bei CONFID.

5.2. Kurzbeschreibung

GHQUAR

Mit diesem Programm können Tabellen mit bis zu 7 Dimensionen bearbeitet werden. Der Nutzer kann in der Steuerdatei den Prozentsatz für die relative Mindestspannweite eines Quaders, der das geheime Feld schützt, festlegen. Nach dem Programmstart erfolgt eine iterative Abarbeitung aller Untertabellen

[37] Gießing, S.; 1999

nach dem Quaderverfahren. Zu jedem sensiblen Tabellenfeld wird ein Quader gesucht.
Die Auswahl des Sperrquaders erfolgt nach gewissen Kriterien:

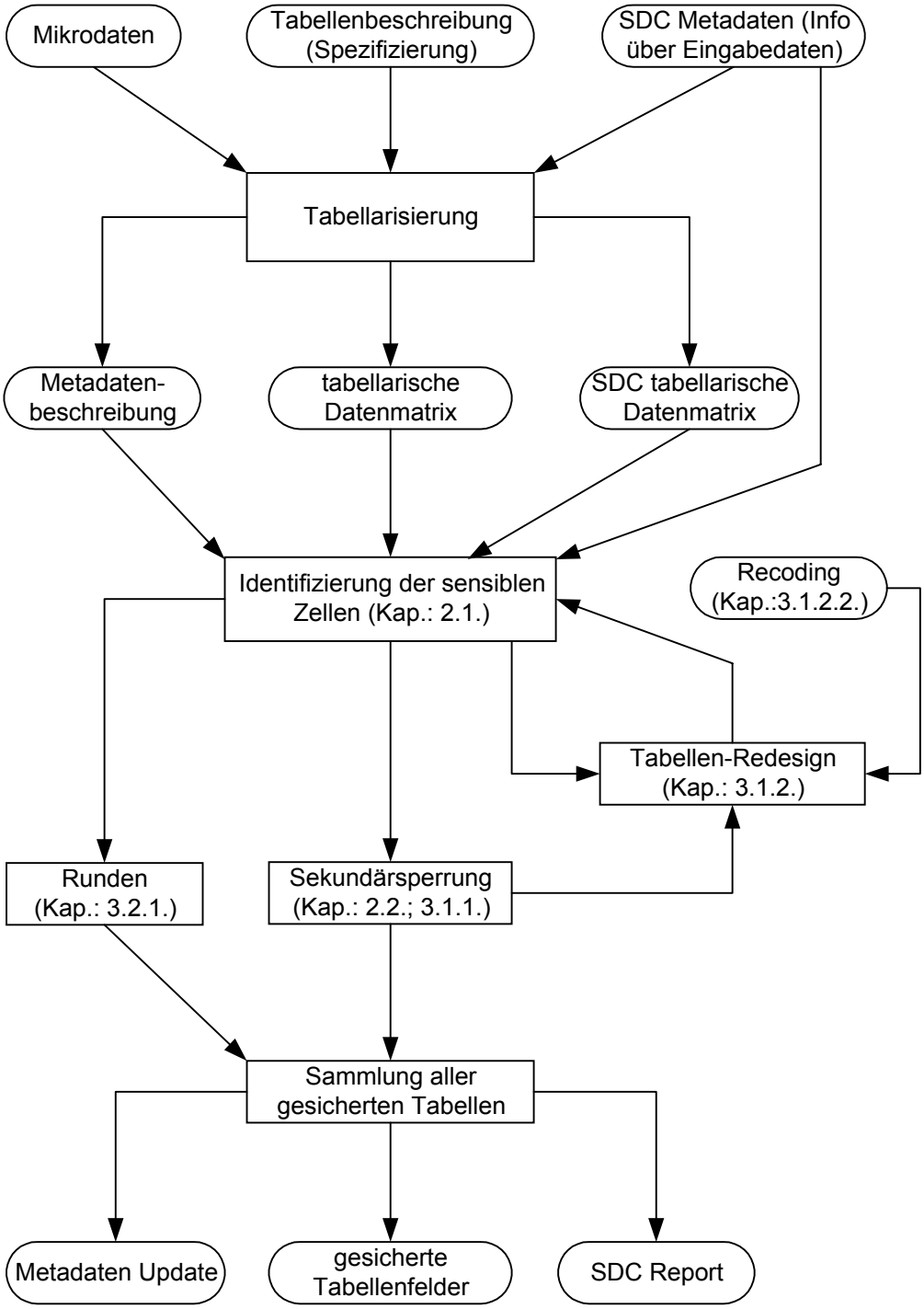
- Minimierung der Anzahl der zu sperrenden Felder
- Minimierung der Wertsumme der gesperrten Felder
- der Prozentsatz der relativen Mindestspannweite muss den vorgegebenen Bedingungen genügen.

Es wird außerdem darauf geachtet, dass möglichst wenig Randsummen gesperrt werden, da bei diesen Werten, der zu minimierende Informationsverlust am größten ist. Die Prozedur wird so oft wiederholt, bis keine Sperrung mehr vorgenommen werden muss. Nach Auskunft des LDS NRW sind dazu durchschnittlich 4 bis 6 Durchläufe erforderlich. Maximal sind 100 Durchläufe möglich.

τ -ARGUS

Mit diesem Programm können maximal vierdimensionale Tabellen (ohne hierarchische Untergliederung durch Zwischensummen) mit Primärsperrungen erzeugt werden. Der Benutzer kann zwischen Sekundärsperrung und kontrolliertem Runden wählen. Nach der Ergebnisanalyse kann noch durch Veränderung des Tabellendesigns (z.B. durch Zusammenfassen von Zeilen und/oder Spalten) versucht werden, die Anzahl der Primärsperrungen und damit die Sperrungen insgesamt zu reduzieren. Sekundärsperrung bzw. Runden kann anschließend erneut durchgeführt werden. Die gesamte Prozedur wird solange wiederholt, bis der Nutzer mit der Menge der gesperrten Felder zufrieden ist. Bei der Sekundärsperrung wird mit einem linearen Optimierungsverfahren gearbeitet, welches das Sekundärsperrungsmuster ermittelt, das den geringsten Informationsverlust sicherstellt.

τ-ARGUS Funktionsdesign[38]:



[38] Hundepool, A.J./ Willenborg, L.; 1998

USBCSUP

Aus den Eingabedaten werden zweidimensionale Untertabellen mit hierarchischem Aufbau konstruiert. Anschließend wird zu jeder Untertabelle ein entsprechendes Netzwerk entwickelt. Die Daten (alle Zellwerte, der jeweils größte und zweitgrößte Einzelbeitrag jeder Zelle und deren Identifikationsschlüssel) der jeweiligen Untertabelle werden aus einer Eingabedatei eingelesen und die darin enthaltenen gesperrten Zellen sukzessive abgearbeitet. Bei der Sperrung weiterer Zellen wird darauf geachtet, dass nur solche Zellkombinationen ausgewählt werden, bei denen das Schutzintervall hinreichend groß ist.

Die Abarbeitung der Untertabellen erfolgt dreimal hintereinander. Im ersten Durchlauf werden bereits früher gesperrte Zellen als mögliche Sperrpartner berücksichtigt. Im zweiten Durchlauf wird ein den Anforderungen des Intervallschutzes genügendes Sperrmuster gebildet. Während des dritten Durchlaufs wird rückwirkend überprüft, ob eventuell ein Teil der gesperrten Zellen wieder freigegeben werden kann. Eine rückwirkende Überprüfung ist nötig, da die verwendete Heuristik zur linearen Optimierung nicht immer die exakte Lösung des Sekundärsperrungsproblems findet. Es ist demnach möglich, dass nicht ganz optimale Ergebnisse geliefert werden, in denen zu viele Tabellenfelder gesperrt wurden.

CONFID

Dieses Programm besteht im Wesentlichen aus zwei Teilmodulen. Mit dem ersten Modul BUILD wird die Struktur der zu verarbeitenden Tabelle eingelesen und aus dem Einzelmaterial eine maximal dreidimensionale Tabelle erstellt. Anschließend wird die Primärsperrung durchgeführt und die Grenzen des Schutzintervalls für jede Primärsperrung ermittelt. Mit dem zweiten Modul SUPRESS werden nun die Gleichungen der linearen Optimierungsaufgabe entwickelt und mit deren Hilfe sukzessive zu jeder Primärsperrung die erforderlichen Sekundärsperrungen mit ausreichendem Intervallschutz ermittelt. Zur weiteren Analyse wird die Routine BOUND eingesetzt. BOUND ermittelt zu jeder gesperrten Zelle die obere und untere Schranke. Diese Routine kann auf jede Tabelle mit Sperrpositionen angewendet werden.

5.3. Konzeptioneller Vergleich[39]

5.3.1. Einsatzmöglichkeiten

Die Einsatzmöglichkeiten hängen von der Struktur, der Gliederungstiefe und dem Umfang der zu behandelnden Tabellen ab.

GHQUAR

Tabellendimension – Mit diesem Programm können bis zu siebendimensionale Tabellen mit hierarchischer Untergliederung durch Zwischensummen bearbeitet werden. Die Sekundärsperrung erfolgt in allen zweidimensionalen Untertabellen.

Rechenzeit – Nach Angabe des LDS NRW werden bei ca. 580 000 Sätzen mit zweidimensionaler Gliederung ca. 4 CPU-Minuten benötigt. In einem durch das Statistische Bundesamt durchgeführten Vergleich der Programme, benötigte GHQUAR für eine dreidimensionale Tabelle mit ca. 50000 Feldern ca. zweieinhalb CPU-Minuten.

Primärsperrungsregel – GHQUAR führt nur die Sekundärsperrung durch. Die Primärsperrung muss vorher erfolgen.

τ-ARGUS

Tabellendimension – Die statistische Geheimhaltung kann bei bis zu vierdimensionalen Tabellen ohne Untergliederungen durch Zwischensummen durchgeführt werden.

Rechenzeit – Es wird davon abgeraten, Sekundärsperrungen für Tabellen mit mehr als 500 Primärsperrungen durchzuführen.

Primärsperrungsregel - τ-ARGUS führt die Primärsperrung mit der (n,k)-Dominanzregel und/oder einer Fallzahlregel durch. Andere Regeln können nicht angewandt werden.

[39] Der gesamte Softwarevergleich wurde durch das Statistische Bundesamt durchgeführt.

USBCSUP

Tabellendimension – Es können nur zweidimensionale Tabellen mit hierarchischer Untergliederung in höchstens einer Dimension in ein Netzwerk umgesetzt werden. Durch Unterteilung höherdimensionaler Tabellen in Teiltabellen, bei denen die Sekundärsperrung iterativ durchgeführt wird, kann das Programm, bei bis zu dreidimensionalen Tabellen, eingesetzt werden.

Rechenzeit – Die Rechenzeiten sind teilweise länger als bei GHQUAR. Für die Abarbeitung der Beispieltabelle (50000 Felder) wurden ca. eineinhalb CPU-Stunden benötigt.

Primärsperrungsregel – USBCSUP bietet die Möglichkeit die Primärsperrung anhand der p%-Regel durchzuführen. Der Benutzer kann jedoch auch bereits festgelegte Primärsperrungsfelder an das Programm übergeben.

CONFID

Tabellendimension – Das Programm kann die statistische Geheimhaltung bei bis zu dreidimensionalen Tabellen mit beliebiger Unterstruktur in jeder Dimension durchführen. Eine iterative Durchführung der Sekundärsperrung für Teiltabellen ist nicht erforderlich.

Rechenzeit – Die Rechenzeiten liegen noch höher als bei USBCSUP. Ab ca. 20000 Tabellenfeldern werden mehrere CPU-Stunden benötigt.

Primärsperrungsregel – Bei aggregierten Tabellenfeldern muss die Primärsperrung mit anderen Programmen durchgeführt werden. Die ermittelten Sperrpositionen werden in der Eingabedatei an das Programm übergeben. Zellsperrungen, die den Wert Null aufweisen, werden ignoriert.

5.3.2. Datensicherheit

Die Programme müssen garantieren, dass die Sekundärsperrung so durchgeführt wird, dass die primär gesperrten Zellen nicht durch einfache Differenzbildung exakt rückrechenbar sind. Bei Sekundärsperrungen, die nur anhand dieses Kriteriums durchgeführt werden, kann in bestimmten Fällen die Datensicherheit gefährdet sein.

GHQUAR

Intervallschutz – Der Benutzer kann den Prozentsatz angeben, der die Größe des Schutzintervalls festlegt. Ausgewählt werden nur Quader, deren Schutzintervall (bezogen auf den größten Quaderwert) größer ist, als die vorgegebene Schutzintervalllänge. Symmetrie ist bei der Wahl des Schutzintervalls nicht erforderlich, d.h. die Primärsperrung muss nicht genau in der Mitte liegen. Es könnten somit Sperrmuster gebildet werden, bei denen der primär gesperrte Wert relativ genau berechnet werden kann, da er mit der oberen oder unteren Schranke übereinstimmt. Dies gilt natürlich nur für höchstens, jeweils eine der beiden Schranken. Bei Anwendung der (n,k)-Dominanzregel sollte die obere Schranke mindestens $\left(\frac{100}{k}\right)\cdot 100\%$ des dominierenden Einzelwertes betragen, um den ungünstigsten Fall (Zellwert=dominierender Einzelwert) abzudecken. Demnach sollte beispielsweise bei Verwendung einer Dominanzregel, mit dem Schwellwert 85, der Prozentsatz für die Größe des Schutzintervalls mit 118% $\left(=\frac{100}{85}\cdot 100\right)$ angegeben werden.

τ-ARGUS

Intervallschutz – Die obere und untere Schranke zum Schutz der primär gesperrten Tabellenfelder wird durch den Anwender festgesetzt. Ähnlich wie bei GHQUAR werden die Prozentsätze des

Zellwertes bestimmt. Bei einer Dominanzschwelle von 85 liegen die Grenzwerte bei 118 bzw. bei 82%.

USBCSUP

Intervallschutz – Das Schutzintervall für eine Primärsperrung wird entsprechend der normierten Sensitivität der gesperrten Zelle festgelegt. Somit ermöglicht weder der obere, noch der untere Grenzwert des Schutzintervalls eine genauere Abschätzung der beitragenden Einzelangaben, als das gemäß der Primärsperrungsregel akzeptiert werden kann.

CONFID

Intervallschutz – Aufgrund des Vorwissens eines Angreifers muss davon ausgegangen werden, dass dieser jede Zelle einer statistischen Tabelle auf 50% genau schätzen kann. Deshalb wird das Schutzintervall für eine Primärsperrung auf ± 50% der normierten Sensitivität der gesperrten Zelle festgesetzt. Es werden nur solche Sperrmuster gebildet, bei denen zu jeder gesperrten Zelle Werte gefunden werden können, die höchstens um 50% und bei primär gesperrten Zellen um mindestens 50% der normierten Sensitivität von den Originalwerten abweichen.

5.3.3. Flexibilität

Es ist notwendig, dass die Programme über gewisse Steuerungsmöglichkeiten verfügen. Beispielsweise gibt das Programm Kriterien für den Informationsverlust vor (viele kleine Zellen oder wenige große Zellen sperren). Jeder Benutzer favorisiert möglicherweise ein anderes Kriterium.

GHQUAR

Sperrmuster – In erster Line wird bei GHQUAR die Zahl der gesperrten Zellen minimiert, erst in zweiter Linie die Summe der gewichteten Zellwerte. Die Gewichtungsvariablen können vom Benutzer frei vorgegeben werden. Des weiteren kann das Programm so

gesteuert werden, dass Sekundärsperrungen in Zwischen- und/oder Randsummen vermieden werden. Schon gesperrte Tabellenfelder werden negativ gewichtet. Dadurch wird sichergestellt das Sperrmuster mit möglichst wenigen zusätzlichen Sperrungen bevorzugt werden.

Intervallausgabe – Die Quaderspannweiten können ausgegeben werden. Die exakten, nur durch lineare Optimierung bestimmbaren Schätzintervalle werden nicht berechnet.

τ-ARGUS

Sperrmuster – Mit den verfügbaren Optionen kann eine Minimierung der gesperrten Wertsumme erreicht werden. Zusätzlich kann der Benutzer eine eigene Gewichtungsvariable für den Informationsverlust definieren.

Intervallausgabe – Die Schätzintervalle können nicht ausgegeben werden. Es besteht jedoch die Möglichkeit einer Programmmodifikation.

USBCSUP

Sperrmuster – In diesem Programm besteht keine Option für die Wahl des Informationsverlustes. Eine entsprechende Programmänderung wäre gegebenenfalls möglich.

Intervallausgabe – In dem vom Statistischen Bundesamt getesteten Programm können die Schätzintervalle nicht ausgegeben werden. Möglicherweise existiert beim US Census Büro ein entsprechendes Zusatzprogramm.

CONFID

Sperrmuster – Der Benutzer kann zwischen zwei verschiedenen Optionen wählen: (a) „size costs“ (Minimierung der gesperrten Wertsumme)
(b) „digit costs“ (Kostenkoeffizient verhält sich in etwa wie der Logarithmus des Zellwertes; diese Option stellt einen Kompromiss zwischen Minimierung der gesperrten Wert-

summe und Minimierung der Anzahl der gesperrten Tabellenfelder dar)

Intervallausgabe – Bei CONFID besteht die Möglichkeit der Schätzintervallausgabe.

5.4. Empirischer Vergleich

5.4.1. Einführung

Alle vorgestellten Programme verfolgen prinzipiell die gleichen Ziele. Die Sekundärsperrung soll so durchgeführt werden, dass die zu den primär gesperrten Tabellenfeldern beitragenden Einzelangaben durch ein hinreichend großes Schutzintervall vor Offenlegung geschützt werden und gleichzeitig soll der Informationsverlust, gemessen an bestimmten Optimalitätskriterien, möglichst gering gehalten werden.
Um zu bestimmen, ob eine Überlegenheit eines der Programme hinsichtlich des Umfangs der Sekundärsperrungen besteht, wurde im StBA ein empirischer Vergleich durchgeführt. Die Sekundärsperrung wurde beispielhaft an einigen Tabellen aus dem Tabellenprogramm der Handwerkszählung 1995 durchgeführt und der Umfang der Sekundärsperrungen in Bezug auf Anzahl und Wert(summe) verglichen. Des weiteren wurde ermittelt, ob größere Unterschiede hinsichtlich der benötigten Rechenzeit bestehen.

5.4.2. Aufbau des Vergleichs

Es wurden sieben Tabellen (Tabelle 1-4: zweidimensional und Tabelle 5-7: dreidimensional) aus dem Tabellenprogramm der Handwerkszählung ausgewählt. Nach Abschluss der Primärsperrung wurde die Sekundärsperrung für diese Tabellen getrennt (keine tabellenübergreifende Geheimhaltung) mit den verschiedenen Programmen durchgeführt.
Mit der von τ-ARGUS vorliegenden Prototypversion konnten die Originaltabellen wegen ihrer komplexen Struktur nicht bearbeitet werden. Jedoch wurde ein

Vergleich zwischen τ-ARGUS und GHQUAR an einer Untertabelle (7a) der Tabelle 7 durchgeführt.
Tabellenfelder mit dem Wert Null wurden weder als Primär- noch als Sekundärsperrposition verwendet, da USBCSUP und CONFID primär gesperrte Nullen ignorieren und andere Tabellenfelder mit dem Wert Null grundsätzlich nicht als Sekundärsperrpartner verwendet.

5.4.3. Ergebnisse

5.4.3.1. Umfang der Sekundärsperrungen

Die folgenden Schaubilder 1 und 2 zeigen Anzahl und Wertsumme der bei den sieben von USBCSUP, vier von CONFID sowie 7a von τ-ARGUS vorgenommen Sekundärsperrungen im Verhältnis zu den Ergebnissen von GHQUAR. Bei Tabelle 1 in Schaubild 2 lagen keine Ergebnisse der von CONFID gesperrten Werte vor.

Abbildung 1: Verhältnis der Anzahl der Sekundärsperrungen zum GHQUAR-Resultat in % je Testtabelle

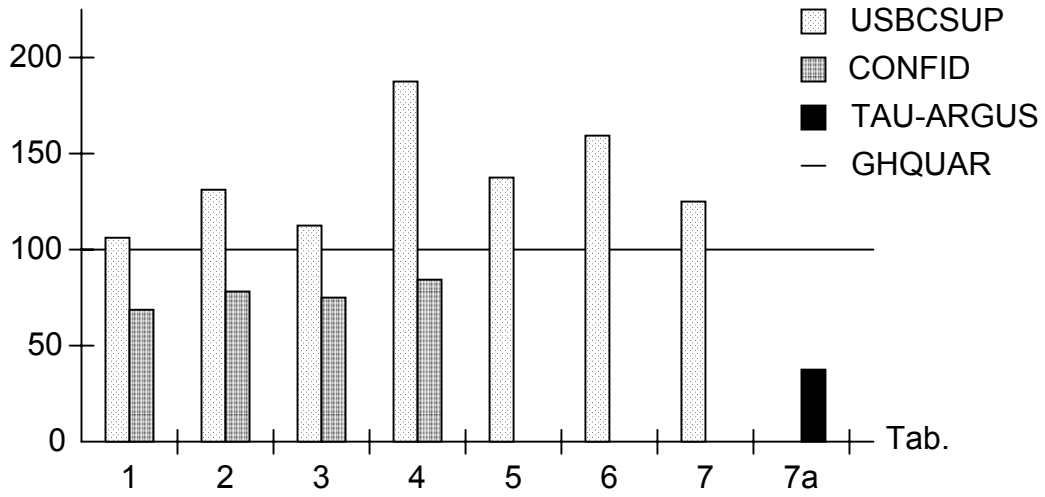
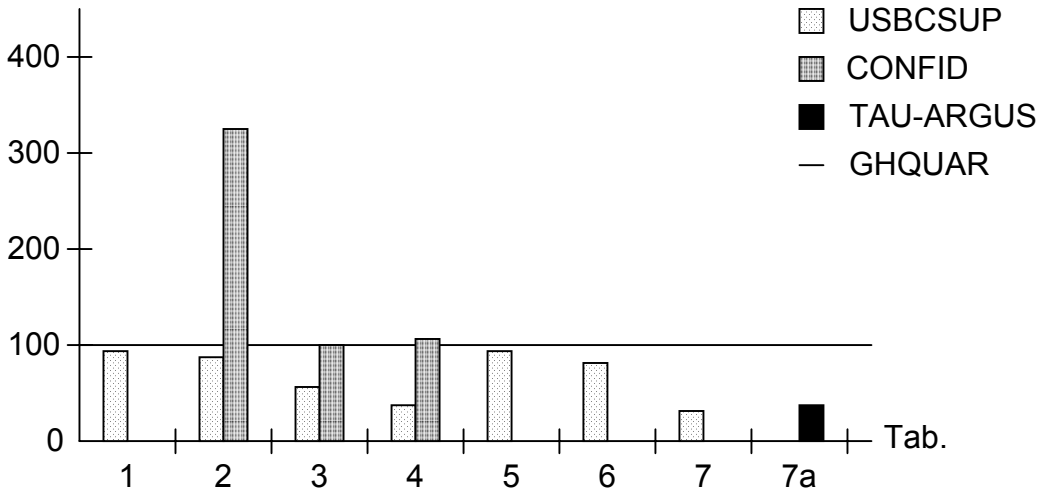


Abbildung 2: Verhältnis der Summe der sekundär gesperrten Werte zum GHQUAR-Resultat in % je Testtabelle



Bei der Anzahl der gewählten Sekundärsperrungen lieferte CONFID bei allen zweidimensionalen Tabellen das günstigste Ergebnis. Bezogen auf die gesperrte Wertsumme war USBCSUP den anderen Programmen überlegen. USBCSUP bevorzugt dabei Sperrmuster bei denen zwar relativ viele, aber überwiegend kleine Zellen gesperrt werden. Hingegen wählt das Quaderverfahren GHQUAR Sperrmuster mit geringer Wertsumme nur dann, wenn dabei nicht mehr Zellen gesperrt werden, als bei alternativen Lösungen mit höherer gesperrter Wertsumme. Bei der Untertabelle 7a lieferte τ-ARGUS die besten Ergebnisse, sowohl in Bezug auf die Anzahl, als auch auf die gesperrte Wertsumme.

5.4.3.2. Rechenzeiten

Die Unterschiede zwischen den Rechenzeiten waren sehr erheblich. GHQUAR war mit Abstand das schnellste, der vier getesteten Programme. Von den anderen Programmen war das mit Netzwerkoptimierung arbeitende Programm USBCSUP deutlich schneller als die mit rechenzeitintensiver linearer Optimierung arbeitenden Systeme CONFID und τ-ARGUS.

Folgende Tabelle gibt einen Ergebnissüberblick:

Ta-belle	Anzahl der Tabellenfelder	Rechenzeit (CPU - Zeit in Stunden:Minuten:Sekunden			
		USBCSUP	CONFID	GHQUAR	Tau-ARGUS
1	6302	00:01:05	00:01:45	00:00:04	#[2)]
2	4630	00:01:20	unter	00:00:03	#[2)]
3	1735	00:00:02	5 Minuten	00:00:01	#[2)]
4	1312	00:01:19		00:00:01	#[2)]
5	47374	01:06:55	#[2)]	00:01:39	#[2)]
6	53045	01:31:44	#[2)]	00:01:34	#[2)]
7	17040	00:03:28	über 4 Std.	00:02:34	#[2)]
7a	2295[1)]	#[2)]	#[2)]	00:00:01	00:21:09

[1)] In der τ-ARGUS Darstellung besteht die Tabelle 7a aus 15283 Feldern (2329 besetzten und 12954 leeren Feldern).

[2)] Lauf nicht durchgeführt oder abgebrochen, wegen extrem langen Rechenzeiten.

5.5. Evaluierung der Ergebnisse und Ausblick

5.5.1. Evaluierung der Ergebnisse

Alle vier getesteten Programme waren in der Lage, die Sekundärsperrung adäquat durchzuführen. Es zeigten sich jedoch erhebliche Qualitätsunterschiede, was Art und Umfang der Tabellen betrifft, die verarbeitet werden können, bei der Rechenzeit, beim Umfang der Sekundärsperrungen und beim Benutzerkomfort.

GHQUAR – arbeitet sehr schnell und ist damit zur Zeit das einzige Programm, das auch bei großen, mehrdimensionalen, durch Zwischensummen hierarchisch gegliederten Tabellen eingesetzt werden kann. GHQUAR schneidet bei der Anzahl der Sekundärsperrungen etwas besser als USBCSUP und in Bezug auf die Wertsumme der Sekundärsperrungen etwas besser als CONFID ab.

USBCSUP – überzeugte im Hinblick auf Benutzerfreundlichkeit und Flexibilität.

Es werden zwar relativ viele Zellen gesperrt, jedoch ist die gesperrte Wertsumme im Mittel verhältnismäßig klein. Die langen Rechenzeiten, die bei der Bearbeitung größerer dreidimensionaler Tabellen anfallen, sind noch vertretbar, wenn nur einzelne Tabellen bearbeitet werden sollen.
CONFID – setzte bei den Testtabellen die wenigsten Sekundärsperrungen. Es kann jedoch wegen des Rechenzeitbedarfs nur bei mäßig großen Tabellen eingesetzt werden. Das Programm ermittelt, die durch die lineare Optimierung berechenbaren Wertebereiche, für alle gesperrten Zellen. Werden diese Wertebereiche veröffentlicht, lässt sich der Informationsverlust stark reduzieren.
τ-ARGUS 1.5[40] – ermittelt optimale Sperrmuster mit minimalem Informationsverlust. Die getestete Version kann allerdings nur bestimmte Typen von Tabellen, die insbesondere keine Zwischensummen aufweisen dürfen, verarbeiten. Somit ist der Einsatz von τ-ARGUS nur bedingt empfehlenswert. Eine Folgeversion für Tabellen mit Zwischensummen war zu dieser Zeit jedoch schon geplant und ist mittlerweile auch erhältlich (τ-ARGUS 2.0).

5.5.2. Empfehlungen und Ausblick

Wegen seiner kurzen Rechenzeiten ist GHQUAR das einzig empfehlenswerte Programm zur Bearbeitung von sehr tief gegliederten und umfangreichen Tabellen. Obwohl CONFID die besten Ergebnisse im Hinblick auf die Zahl der Sekundärsperrungen erzielt, kann der Einsatz nicht empfohlen werden. Gründe hierfür sind einerseits die relativ hohen Kosten und andererseits die beschränkte Einsatzfähigkeit hinsichtlich der Tabellengröße. Die getestete τ-ARGUS Version ist für den Einsatz in der amtlichen deutschen Statistik ungeeignet. Der Sekundärsperrungsalgorithmus von τ-ARGUS verspricht eine hervorragende Ergebnisqualität hinsichtlich des Informationsverlusts. Mit einer stetigen Weiterentwicklung des Programms sind derzeit das StBA, das CBS Niederlande sowie Forscher verschiedener europäischer Universitäten beschäftigt. Es ist vorgesehen, das Quaderverfahren als einen von mehreren alternativen Sekundärsperrungsalgorithmen in das Programm einzubinden.

[40] de Jong, W.; 1992

6.Schlussbemerkungen

6.1.Zusammenfassung

Daten, welche zu statistischen Zwecken aufgearbeitet werden, sind öffentliche Güter und somit jedem zugänglich zu machen. Da schon allein aus gesetzlichen Vorschriften die Notwendigkeit zur vertraulichen Behandlung von Einzelangaben besteht, sind die vorgestellten Verfahren und Methoden von elementarer Bedeutung. Die Veröffentlichung aggregierter Unternehmensdaten stellt kein datenschutzrechtliches Problem dar, sofern in den Tabellen eine genügende Anzahl von Untersuchungseinheiten zusammengefasst sind und nur Durchschnitte oder Summen ausgewiesen werden. Für diese Daten bestehen lediglich die Probleme:

- einer nutzerfreundlichen Information über die verfügbaren aggregierten Daten und Zeitreihen,
- der Standardisierung der Konzepte, der korrekten Erfassung und der Vergleichbarkeit in der zeitlichen, regionalen und internationalen Dimension,
- der verlässlichen (regelgebundenen) und schnellen Zugänglichkeit am Arbeitsplatz des Benutzers (dazu gehören auch Regeln, wie und wann solche Statistiken veröffentlicht werden)

Gesamtwirtschaftlich sind diese Tabellen für die Benutzer von großer Bedeutung. Sie bieten Planungssicherheit, zeigen Trends und Entwicklungstendenzen auf und informieren über die allgemeine Wettbewerbs- und Wirtschaftssituation. Da jedoch alle Unternehmen innerhalb eines geschlossenen Systems den Zustand der vollständigen Information anstreben, ist die Geheimhaltungspflicht auch für negative Effekte verantwortlich. Wie bei der Vorstellung der Verfahren schon festgestellt wurde, geht jeder Geheimhaltungsalgorithmus mit einem unerwünschten Informationsverlust einher. Da jegliches Informationsdefizit auch Wettbewerbsnachteile und Planungsunsicherheit bedeutet, ist es nicht auszuschließen, dass die Benutzer andere Wege finden, die sicheren Daten

aufzudecken. Auswirkungen wären Kostenexplosion, Wirtschaftskriminalität, Monopol- und Kartellbildung, d.h. Absprachen zu Lasten Dritter.

Folgendes Szenario wäre denkbar: In einem geschlossenen Wirtschaftsraum mit einer begrenzten Anzahl an Wettbewerbern (z.B. 3 Bäcker), könnte Bäcker A durch Absprache mit Bäcker B geheime Daten des Bäcker C aufdecken. Durch abgestimmte Handlungsweisen von Bäcker A und B könnte Bäcker C vom Markt gedrängt werden. Somit wäre, motiviert aus dem Drang nach vollkommener Information, der Wettbewerb eingeschränkt und die Bildung eines Preiskartells erfolgt. Ähnlich verhält es sich im umgekehrten Fall. Angenommen es gäbe nur einen Bäcker A. Dann würden seine Einzelangaben nicht veröffentlicht werden (siehe Primärsperrung; Fallzahl=1). Er besitzt also im abgegrenzten System eine Monopolstellung. Möglicherweise wird dies durch den Zustand der völligen Ignoranz (Informationsgrad nahe Null) auch so bleiben, weil der Eintritt für Dritte in diesen Markt mit zu vielen Risiken verbunden ist.

Trotz der aufgezeigten Nachteile, welche durch Geheimhaltungsbestimmungen hervorgerufen werden könnten, ist wohl unbestritten die Notwendigkeit zum Schutz des Einzelnen zu erkennen.

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.

6.2.Ausblick

Die Gewährleistung der statistischen Geheimhaltung ist eine hoheitliche Aufgabe in der amtlichen Statistik. Sie gründet sich unter anderem auf das Bundesstatistikgesetz. Die sekundäre Geheimhaltung sichert in aggregierten Daten bereits gesperrte, d.h. primär geheime Tabellenwerte gegen zu genaue Rückrechnung durch Unterdrückung zusätzlicher Tabellenwerte, die Sekundärsperrungen. Um den damit verbundenen Informationsverlust klein zu halten, wird eine möglichst kleine Anzahl von Sekundärsperrungen und eine kleine Summe zu sperrender Werte angestrebt. Das Geheimhaltungsproblem wird von den statistischen Ämtern vieler Länder bearbeitet, was bisher zu recht unterschiedlichen Verfahren geführt hat. Das im LDS NRW entwickelte Quaderverfahren hat sich in dem vom Statistischen Bundesamt vorgenommenen weltweiten Vergleich als das bisher effizienteste Verfahren erwiesen. Neben den Untersuchungen des Statistischen Bundesamtes hat nun auch EUROSTAT Vergleiche mit einem für das Quaderverfahren ähnlich günstigen Ergebnis durchgeführt.

EUROSTAT hat daher eine externe Firma mit der Erstellung eines nutzerfreundlichen Interfaces (CIF) zur Steuerung des auf dem Quaderverfahren basierenden EDV-Programms GHQUAR beauftragt. Im Rahmen dieser Arbeiten wurde GHQUAR auf PC umgestellt. Das Interface CIF wie auch das Geheimhaltungsprogramm GHQUAR können nunmehr auf allen gängigen Betriebssystemen eingesetzt werden. Das von der externen Firma fertiggestellte Interface ist für die Sicherung von einzelnen Statistiktabellen konzipiert worden. Es kann jedoch keine EDV-Programme steuern, die auch mehrere einander überlappende Tabellen bearbeiten, wie dies z.B. mit dem EDV-Programm GHMITER (Geheimhaltung mit iterativem Einzeltabellenabgleich), einer Erweiterung von GHQUAR, möglich ist. Aus diesem Grunde wurde nun von EUROSTAT ein Fortsetzungsvertrag für die Erweiterung des CIF-Interfaces mit der externen Firma vorbereitet. Mit der neuesten Version von CIF und GHMITER können bis zu 20 Einzeltabellen durch Übertragung der Schutzintervalle der mehreren Tabellen gemeinsam angehörenden geheimen Werte gesichert werden.

Literaturverzechnis

[1] **Adam, N.R./ Wortmann, J.C.:** Security Control Methods for Statistical Databases. In: ACM Computing Surveys, vol. 21, S. 521, 1989

[2] **Als, G.:** Propositions on Statistical Confidentiality. In: International Seminar on Statistical Confidentiality in Luxembourg 1994

[3] **Block, H.:** Confidentiality at the Statistics Sweden. In: International Seminar on Statistical Confidentiality in Dublin 1992

[4] **Causey, B.D./ Cox, L.H./ Ernst, L.R.:** Applications of Transportation Theory to Statistical Problems. In: Journal of the American Statistical Association, S. 903-909, 1985

[5] **Cox, L.H.:** Suppression Methodology and Statistical Disclosure Control. In: Journal of the American Statistical Association, Vol. 75, No 370, U.S. Bureau of the Census 1980

[6] **Cox, L.H.:** Linear Sensitivity Measures in Statistical Disclosure Control. In: Journal of Planning and Inference, 5, 153 - 164, 1981

[7] **Cox, L.H./ Ernst, L.R.:** Controlled Rounding. INFOR, S. 423-432, 1982

[8] **Cox, L.H.:** Some Mathematical Problems Arising from Confidentiality Concerns. In: Special Issue of the Statistical Review dedicated to T. Dalenius, S.179-189, 1983

[9] **Cox/ Fagan/ Greenberg und Hemmig:** Research at the Census Bureau into Disclosure Avoidance Techniques for Tabular Data. In: Journal of the American Statistical Association, S. 388-393, 1986

[10] **Cox, L.H.:** A Constructive Procedure for Unbiased Controlled Rounding. In: Journal of the American Statistical Association, S. 520-524, 1987

[11] **Cox, L.H./ George, J.A.:** Controlled Rounding for Tables with Subtotals. In: Annals of Operation Research 20, S. 141-157, 1989

[12] **Cox, L.H.:** Solving Confidentiality Protection Problems in Tabulations using Network Optimization: A Network Model for Cell Suppression. In: International Seminar on Statistical Confidentiality in Dublin 1992

[13] **Dalenius, T.:** Towards a Methodology for Statistical Disclosure Control. In: Statistisk Tidskrift, S. 429-444, 1977

[14] **Dalenius, T.:** A Simple Procedure for Controlled Rounding. In: Norstedts Tryckeri, Stockholm, 1981

[15] **Dalenius, T.:** Controlling Invasion of Privacy in Surveys. In: Statistical Research Unit, Statistics Schweden, 1988

[16] **Dellaert, N.P./ Luijten, W.A.:** Statistical Disclosure in General Three-dimensional Tables. In: Statistica Neerlandica, S. 197-221, 1999

[17] **Domingo-Ferrer, J./ Mateo-Sanz, J.M.:** On Resampling for Statistical Confidentiality in Contingency Tables. In: Report, Universitat Rovira I Virgili, Tarragona, Spanien, 1999

[18] **Duarte de Carvalho, F.:** Statistical Disclosure in Two-dimensional Tables. In: Journal of the American Statistical Association, S. 1547-1557, 1994

[19] **EUROSTAT:** Manual on Disclosure Control Methods. Luxemburg 1996

[20] **Evans, B.T./ Zayatz, L.:** Using Noise for Disclosure Limitation of Establisment Tabular Data. In: Proceedings of the 1996 Annual Research Conference, U.S. Bureau of the Census, S. 65-86, 1996

[21] **Fellegi, I.P.:** Controlled Random Rounding. In: Survey Methodology, S. 123-133, 1975

[22] **Fischetti, M./ Salazar, J.J.:** Computational Experience with the Controlled Rounding Problem. In: Report, University of La Laguna Tenerife, 1996

[23] **Fischetti, M./ Salazar, J.J.:** Models and Algorithms for the Cell Suppression Problem. In: International Seminar on Statistical Confidentiality in Bled 1996

[24] **Fischetti, M./ Salazar, J.J.:** Models and Algorithms for Optimizing Cell Suppression in Tabular Data. In: Report, University of La Laguna Tenerife, 1998

[25] **Fischetti, M./ Salazar, J.J.:** Modelling and Solving the Cell Suppression Problem for Linearly-Constrained Tabular Data. In: SDP Conference in Lissabon 1998

[26] **George, J.A./ Penny, R.:** Preservation of Confidentiality in Aggregated Data. In: International Seminar on Statistical Confidentiality in Bled 1996

[27] **Geurts, J.:** Netherlands Central Bureau of Statistics, Department of Statistical Methods, P.O. Box 959, 2270 AZ Voorburg, Heuristics for Cell Suppression in Tables, 1992

[28] **Giessing, S.:** Looking for efficient automated secondary cell suppression systems: a software comparison. In: Official Statistics Journal 2/98

[29] **Giessing, S.:** A survey on packages for automated secondary cell suppression. Statistisches Bundesamt. In: Eurostat/UN-ECE Work Session on Statistical Data Confidentiality in Thessaloniki 1999

[30] **Giessing, S.:** Vergleich der Software zur maschinellen Durchführung der Sekundären Geheimhaltung. In: Forum der Bundesstatistik, Band 31/1999: Methoden zur Sicherung der Statistischen Geheimhaltung 1999

[31] **Giessing, S.:** Statistische Geheimhaltung in Tabellen. In: Forum der Bundesstatistik, Band 31/1999: Methoden zur Sicherung der Statistischen Geheimhaltung 1999

[32] **Giessing, S.:** New tools for cell suppression in Tau-ARGUS**.** In: proceedings of the Eurostat/UN-ECE Work Session on Statistical Data Confidentiality 2001

[33] **Giessing, S.:** The CASC Project: Integrating Best Practice Methods for Statistical Confidentiality. In: proceedings of the NTTS & ETK Conference 2001.

[34] **Greenberg, B.:** Disclosure Avoidance Research at the Census Bureau. In: Proceedings of the Annual Research Conference, Washington, 1990

[35] **Herr, G.:** A Bootstrap Procedure to Preserve Statistical Confidentiality in Contingency Tables. In: International Seminar on Statistical Confidentiality in Dublin, S. 261-271, 1992

[36] **Hundepool, A.J./ Willenborg, L.:** Tau-Argus (Version 2) User's Manual. In: Report, Statistics Netherlands, Voorburg, 1998

[37] **Jewett, R.:** Disclosure Analysis for the Economic Census. In: Economic Statistical Methods and Programming Division, Bureau of the Census, Washington, DC 1993

[38] **de Jong, W.:** ARGUS – An Integrated System for Data Protection. In: International Seminar on Statistical Confidentiality in Dublin, S. 317-322, 1992

[39] **Kelly, J.P./ Golden, B.L./ Assad, A.A.:** Using Simulated Annealing to Solve Controlled Rounding Problems. In: Operations Research Society of America Journal on Computing, S. 174-185, 1990

[40] **Kelly, J.P./ Golden, B.L./ Assad, A.A.:** Cell Suppression Using Sliding Protection Ranges. In: Working Paper Series MS/S 90-007, University of Maryland, 1990

[41] **Kelly, J.P./ Golden, B.L./ Assad, A.A.:** Cell Suppression: Disclosure Protection for Sensitive Tabular Data. Networks, *22, S.* 397– 417, 1992

[42] **Kirkendall, N./ Sande, G.:** Comparison of Systems Implementing Automated Cell Suppression for Economic Statistics. In: Journal of Official Statistics, S. 513-535, 1998

[43] **Kumar, R./ Golden, B.L./ Assad, A.A.:** Cell Suppression Strategies for Three-Dimensional Tabular Data. In: Roceedings Annual Research Conference, S. 77-104, Bureau of the Census USA; 1992

[44] **Lougee-Heimer, R.:** Guaranteeing Confidentiality: The Protection of Tabular Data. In: Master's Thesis, Department of Mathematical Sciences, Clemson University, 1989

[45] **Massel, P.B.:** Cell Suppresion and Audit Programs used for Economic Magnitude Data. In: SRD Research Report Series No. RR2001/01, Bureau of the Census, Washington D.C., 2001

[46] **Nargundkar, M.S./ Saveland, W.:** Random Rounding: A Means of Preventing Disclosure of Information about Individual Respondents in Aggregate Data. In: Report, Statistics Canada, Ottawa, 1972

[47] **Repsilber, R.D.:** EDV-Verfahren zur Wahrung der Geheimhaltung bei Tabellen mit bis zu sieben Ordnungskriterien. In: Statistische Rundschau Nordrhein-Westfalen 1991

[48] **Repsilber, R.D.:** Landesamt für Datenverarbeitung und Statistik Nordrhein-Westfalen, Safeguarding Secrecy in Aggregative Data. In: International Seminar on Statistical Confidentiality in Dublin 1992

[49] **Repsilber, R.D.:** Preservation of Confidentiality in Aggregated Data. In: International Seminar on Statistical Confidentiality in Luxembourg 1994

[50] **Repsilber, R.D.:** Wahrung der Geheimhaltung in aggregierten Daten Quaderverfahren mit Intervallschutz für vollständige Tabellen. In: Forum der Bundesstatistik, Bd. 31/1999, Methoden zur Sicherung der Statistischen Geheimhaltung 1999

[51] **Robertson, D.:** The general problem of disclosure avoidante. Statistics Canada. Internal Report 1991

[52] **Robertson, D.:** Cell Suppression at Statistics Canada. In: Proceedings Annual Research Conference, U.S. Bureau of the Census 1993

[53] **Robertson, D.:** Automated Disclosure Control at Statistics Canada. In: International Seminar on Statistical Confidentiality in Luxemburg 1994

[54] **Robertson, D.:** Improving Statistics Canada's cell suppression software (CONFID). In: Proceedings of the Compstat 2000 Conference, Utrecht, Netherlands, 2000

[55] **Sande, G.:** Structure of the Automated Cell Suppression System. In: Eurostat/UN-ECE Work Session on Statistical Data Confidentiality in Thessaloniki 1999

[56] **Sande, G.:** Automated Cell Suppression to Preserve Confidentiality of Business Statistics. In: Statistical Journal of the United Nations, ECE 2, S. 33-41, 1984

[57] **de Vries, R.E.:** Disclosure Control of Tabular Data Using Subtables. In: Report, Department of Statistical Methods, Statistics Netherlands, Voorburg, 1993

[58] **de Waal, A.G.:** The Number of Tables to be Examined in a Disclosure Control Procedure. In: Report, Department of Statistical Methods, Statistics Netherlands, Voorburg, 1993

[59] **Willenborg, L./ de Waal, A.G.:** Optimum Global Recoding and Local Suppression. In: Report, Department of Statistical Methods, Statistics Netherlands, Voorburg, 1995

[60] **Willenborg, L./ de Waal, T.:** Statistical Disclosure Control in Practice. In: Lecture Notes in Statistics 111, Springer Verlag, New York 1996

[61] **Willenborg, L./ de Waal, T.:** Exact Disclosure in a Super-Table. In: Netherlands Official Statistics, S. 11-16, 1998

[62] **Willenborg, L./ de Waal, T.:** Elements of Statistical Disclosure Control. In: Lecture Notes in Statistics 155, Springer Verlag, New York 2001

[63] **Willenborg, L./ Hundepool, A.:** ARGUS: Software from the SDC Project. In: Eurostat/UN-ECE Work Session on Statistical Data Confidentiality in Thessaloniki, S. 87-98, 1999

[64] **de Wolf, P.P.:** A Heuristic Approach to Cell Suppression in Hierarchical Tables. In: Department of Statistical Methods, Voorburg, 1999

[65] **Zayatz, L.:** U.S. Bureau of the Census, Using Linear Programming Methodology for Disclosure Avoidance Purposes. In: International Seminar on Statistical Confidentiality in Dublin 1992

[66] **Zayatz, L./ Massell, P./ Steel, Ph.:** Disclosure Limitation Practices and Research at the U.S. Census Bureau. In: Netherlands Official Statistics, S. 26-29, 1998

Internetadressen

http://europa.eu.int/comm/eurostat
http://neon.vb.cbs.nl/casc
http://www.cbs.nl
http://www.census.gov
http://www.destatis.de
http://www.lds.nrw.de
http://www.oestat.gv.at
http://www.statcan.ca
http://www.statistik.saarland.de
http://www.unece.otg/stats

Ehrenwörtliche Erklärung

Ich versichere, dass ich die Arbeit selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen oder anderen Quellen entnommen sind, sind als solche kenntlich gemacht.

Ilmenau, den 08.02.2002

Unterschrift