aber in keinem Fall die Unabhängigkeit und Weisungsfreiheit des Datenschutzbeauftragten beeinträchtigen.

7 Zusammenfassung

Im Ergebnis ist festzustellen, daß die Prüfung der ordnungsgemäßen Programmanwendung weitestgehend durch den Datenschutzbeauftragten selber durchgeführt werden sollte. Aufgrund der bereits im Unternehmen vorhandenen Richtlinien und Arbeitsanweisungen einerseits und der immer umfangreicher werdenden Standardisierung und Normierung im Programmierbereich ist es relativ leicht möglich, je nach den organisatorischen und technischen Verhältnissen im Unternehmen, die für die Prüfung der Ordnungsmäßigkeit der Programmanwendung notwendigen Kriterien herauszuarbeiten. Wenn auch Wirtschaftsprüfer und interne Revision im Rahmen ihrer Prüfungsaufgaben Belange des Datenschutzes prüfen, ändert dies nichts an der vom Gesetzgeber gewollten Doppelgleisigkeit der Kontrolle. Die Existenz des betrieblichen Datenschutzbeauftragten und dessen Prüfungstätigkeit, entlasten auch nicht die Leiter der Fachabteilungen von ihren eigentlichen eigenen Kontrollaufgaben. Einer sich möglichen ergebenden Überlappung der Prüfungsaktivitäten zwischen Datenschutzbeauftragten und interner Revision, sollte durch Absprache und gegenseitigen Austausch von Erkenntnissen und Ergebnissen entgegengetreten werden. Auch wenn es die Organisation und die Größe des Unternehmens erforderlich macht, daß der Datenschutzbeauftragte Teilbereiche seiner Prüfungsaktivitäten deligiert, entbindet ihn dies nicht von seiner eigentlichen Verantwortung für diesen vom Bundesdatenschutzgesetz vorgeschriebenen Kontrollauftrag. Im anderen Fall können jedoch sinnvolle Abstimmungen und Abgrenzungen mit interner Revision und Wirtschaftsprüfung verhindern, daß es zu wirtschaftlich nicht vertretbaren Prüfungsredundanzen kommt. Sollten dagegen aus dem Nebeneinander von Datenschutzprüfung durch den Beauftragten und Prüfungsaktivität der Revision in Einzelbereichen Überschneidungen entstehen, kann dies durchaus der vom Gesetzgeber gewollten besonderen doppelten Kontrollfunktion entsprechen.

Anmerkungen

1) Auernhammer, Bundesdatenschutzgesetz, 2. Auflage, Anm. 1 zu § 29, S. 287, 288.
2) Schaffland/Wilffang, Bundesdatenschutzgesetz, Anm. 5 zu § 29, so auch Auernhammer, Anm. 1 zu § 29, S. 288.
3) Ordemann-Schomerus, Bundesdatenschutzgesetz 3. Auflage, Anm. 1 zu § 29, S. 254 und Simitis, Bundesdatenschutzgesetz, Anm. 8 zu § 29, S. 887.
4) Ordemann-Schomerus, Bundesdatenschutzgesetz, 3. Auflage, Anm. 1, 2, 3 zu § 29, S. 256.
5) Simitis/Dammann/Mallmann/Reh, Bundesdatenschutzgesetz 3. Auflage, Anm. 49 ff. zu § 29, S. 901.
6) Simitis, ebenda, S. 901 und Langer, Handbuch des Datenschutzes 1979/80, S. 48; Gallwas, Schneider, Schwappach, Scheinoch, Steinbrink, ebenda Anm. 18 zu § 29.
7) Simitis, ebenda, Anm. 41, S. 901.
8) Heilmann/Wolf, Handbuch des Datenchutzes 1979/80, S. 22; Schwappach, BDSG, Anm. zu § 29; Geiger, DSWR 1978, S. 118.
9) Fachausschuß für moderne Abrechnugnssysteme, (FAMA 1/74, Prüfung von EDV-Buchführung) und FAMA 1/79 zur Auslegung der Grundsätze ordnungsmäßiger Buchführung beim Einsatz von EDV-Anlagen im Rechnungswesen.
10) Wolf in Heilmann, Handbuch
11) FAMA 1/74, 1/75, 1/78, vgl. auch TZ 1. Grundsätze ordnungsmäßiger Speicherbuchführung, BFM Erlaß vom 5.7.1978, BStBl 1978 I, S. 250 ff.
12) Merkblatt für die Prüfung von Rechnungswerken die mit ADV erstellt sind, 1980, Bundesminister der Finanzen.
13) Schuppenhauer, Grundsätze für eine ordnungsgemäße Datenverarbeitung IDW-Verlag, Düsseldorf 1982, S. 29.
14) FAMA 1/79, Bundesdatenschutzgesetz und Jahresabschlußprüfung IDW-Verlag, Düsseldorf 1982, S. 29 ff.
15) Revision der elektronischen Datenverarbeitung, IIR Schriftenreihe, Bd. 1, 4. Auflage 1982, S. 11 ff.
16) Ebenda, Merkblatt für die Prüfung von Rechnungswerken, TZ. 4.2.2.4, S. 5.
17) Fachgutachten 1/77, Grundsätze ordnungsmäßiger Durchführung von Jahresabschlußprüfungen.
18) Datenschutz-Berater, Heft 9/1982, ohne Verfasser.
19) Grünwald, Kühn, Zeitschrift für Interne Revision, 1/1979, S. 28.
20) Simitis, Anmerkung 42 zu § 29, S. 902.
21) Simitis, ebenda, Anm. 16, 17 zu § 29, S. 891 und Randnr. 10, 17 ff. zu § 45.
22) Hinweis Nr. 6 des Badenwürttembergischen Innenministeriums.

Wie anonym sind anonymisierte Datensätze wirklich?

Matthias Rauterberg

Spätestens seit dem Volkszählungsurteil des Bundesverfassungsgerichts rückt die Anonymisierung personenbezogener Daten vom Rande zum Zentrum der Datenschutzdiskussion. Wann darf z. B. das Statistische Bundesamt Daten weitergeben, die es nach Rechtsnorm für die Weitergabe (wofür sonst?) erheben muß. Anonymisierte Daten bieten in dieser Beziehung scheinbar kein Problem, denn bei ihnen ist ja der Personenbezug getilgt; anonymisiert sind sie, wenn die Bezugsperson nicht bestimmbar ist. Und wann ist sie nicht bestimmbar; wenn man die Bezugsperson nicht eindeutig bestimmen aber noch erraten kann; wenn auch dies nicht gelingt? Wenn z. B. an soziodemographischen Daten nur Beruf und Geburtsort geführt werden, wird man eine in Berlin geborene Sekretärin gewiß nicht identifizieren können, wohl aber einen in Ebingen geborenen Altbundeskanzler. Auch wenn man den Geburtsort nicht kennt, wird die Chance, den gemeinten Altbundeskanzler zu erraten mindestens ein Drittel sein. Vollständig anonymisiert kann ein Datensatz nicht sein, wenn „nichtbestimmbar" auch das mehr oder minder leicht „erratbar" ausschließen soll.

Der Leser kann anhand der beiden folgenden Beiträge die Schwierigkeit des Themas geradezu in Augenschein nehmen. Sie bieten sich

dem heute davon betroffenen Statistiker und Wissenschaftler, der einen Konflikt mit den Datenschutzbestimmungen gern ausschließen möchte, auch wenn er weiß, daß weniger heiß gegessen als gekocht wird.

Der Verfasser des ersten Beitrags geht davon aus, daß man nicht für den Einzelfall, sondern nur insgesamt abschätzen kann, wie objektiv sicher anonymisiert eine Datei ist. Er gibt einige Abschätzungsverfahren für geplagte Erhebungen an und vergleicht sie anhand von Beispielen. Sie liefern bezogen auf die geprüfte Datei einen Wert, der oberhalb oder unterhalb einer Schranke liegen kann und damit die Datei als ausreichend oder unzureichend anonymisiert ausweist.

Die Herausgeber

Zusammenfassung

Zur Gewährleistung des Datenschutzes von personenbezogenen Datenbeständen wurde und wird es bisher als hinreichend angesehen, daß die Daten zur direkten Personenidentifikation von der Weiterverwendung ausgeschlossen werden. In wie weit die Anonymität des Datenbestandes an Hand der darüber hinaus erfaßten soziodemographischen, personenbezogenen Daten gewährleistet blieb, ließ sich bisher nur intuitiv im vorhinein abschätzen. Es wird eine Berechnungsvorschrift vorgestellt, die den prozentualen Anteil an de-anonymisierbaren Personen eines als anonym geltenden Datenbestandes an Hand der schon bekannten Verteilungen der soziodemographischen Daten vor der Erhebung abschätzt.

Problemstellung

In weiten Bereichen von Verwaltung, Wirtschaft und Wissenschaft sind die Erstellung von größeren Datenbeständen als Grundlage von Planungs- und Entscheidungsprozessen häufig wünschenswert. In denjenigen Bereichen, in denen die Ansammlung von personenbezogenen Daten als sinnvoll erscheinen, stellt sich das Problem des Datenschutzes und der Datensicherung. Im weiteren Verlauf werde ich das Problem der Deanonymisierbarkeit von Datenbeständen behandeln. Dabei werde ich unter denjenigen Daten, die der *direkten Personenidentifikation* dienen, folgende verstehen:

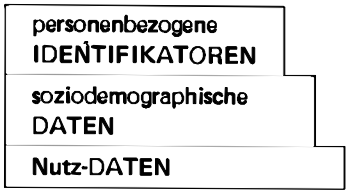
Name, Vorname
Anschrift
Personalausweis-Nummer
etc.

Darüber hinaus werden in der Regel noch *soziodemographische Daten* (SDD) erfaßt, die der weiteren Beschreibung der einzelnen Personen dienen:

Geschlecht
Alter
Staatsangehörigkeit
Schulbildung
Familienstand
Religionszugehörigkeit
etc.

Die bisher genannten Daten zähle ich zu den ‚allgemein' bekannten Daten. Des weiteren werden Daten zu den verschiedenen Verwendungszwecken erhoben. Diese spezielleren Daten zähle ich zu den ‚allgemein' unbekannten Daten und nenne sie auch Nutz-Daten, derentwegen die Erhebung in der Regel durchgeführt werden soll.

Der zu erhebende Datenbestand läßt sich also in drei unterschiedlich große und disjunkte Bereiche aufteilen:



Die gängige Praxis, einen derartigen Datenbestand im Nachhinein vor unberechtigter Re-individualisierung zu schützen, besteht darin, den obersten Teil (die personenbezogenen Identifikatoren) des Datenbestandes vor Weitergabe an Dritte zu entfernen.

Da die Nutz-Daten in der Regel nicht bekannt sind, können sie auch nicht zu einer Re-individualisierung mißbraucht werden. Falls einige doch als bekannt vorauszusetzen sind, können diese Daten unter den SDD eingereiht werden.

Das Problem der im vornhinein abzuschätzenden Anonymität stellt sich nun folgendermaßen:

„Wieviele Personen können an Hand der SDD deanonymisiert werden?"

Um auf diese Frage eine Antwort geben zu können, möchte ich sie so umformulieren, daß sich dann zwanglos der Weg zu einer Berechnungsvorschrift auftut.

Zunächst werde ich den Bereich der SDD präziser darstellen. Es sollen verschiedene SDD erhoben werden, deren Anzahl durch m dargestellt wird.

Jede dieser m verschiedenen SDD hat n_i Ausprägungsstufen (mit $i = 1 \ldots m$). So hat z. B. die soziodemographische Frage (SDF) zum Geschlecht zwei Ausprägungsstufen:

1 = weiblich; 2 = männlich.

Zu jeder Ausprägungsstufe gehört eine aus den allgemeinen Statistiken zu entnehmende, von daher bekannte relative Häufigkeit des Auftretens der einzelnen Ausprägungsstufen in der zu erfassenden Population. Diese relativen Häufigkeiten kürze ich mit p_{i_j} mit $j = 1 \ldots n_i$ und mit $i = 1 \ldots m$ ab. Dabei gilt:

$$\sum_{j=1}^{n_i} p_{i_j} = 1 \quad \text{für alle } i = 1 \ldots m \qquad (1)$$

Diese m verschiedenen SDD spannen dann eine m-dimensionale Matrix auf.

Die Anzahl der Zellen dieser m-dimensionalen Matrix ergibt sich nach folgender Formel:

$$ZZ = \prod_{i=1}^{m} n_i \quad \text{mit ZZ als Zellenzahl.} \qquad (2)$$

In den einzelnen Zellen der SDD stehen nun die absoluten Häufigkeiten derjenigen Personen, auf die das jeweilige Muster, bzw. Merkmalskombination in den Angaben zu den SDD gerade zutrifft.

Wenn in *allen* Zellen *mehr als eine* Person auftaucht, so zähle ich diesen Datenbestand zu den weniger stark gefährdeten Datenbeständen und nenne ihn vorläufig ‚anonym'.

Wenn jedoch in einigen Zellen *nur eine* Person hineingehört, so läßt sich diese Person an Hand ihrer SDD *eindeutig* deanonymisieren.

Nun lautet die umformulierte Frage nach dem Ausmaß der Deanonymisierbarkeit eines Datenbestandes wie folgt:

> „Wieviele einfach besetzte Zellen (EZ) werden in dem zu erhebenden Datenbestand zu erwarten sein?"

Der prozentuale Anteil der eindeutig deanonymisierbaren Personen (DA) errechnet sich nach folgender Formel:

$$DA = \begin{cases} (1.0 * 100)\ \% & ;\text{für } EZ/N > 1 \\ (EZ/N * 100)\ \%; & \text{sonst} \end{cases} \qquad (3)$$

Hierbei bedeutet N die Anzahl der insgesamt gespeicherten Personen.

An dieser Stelle läßt sich schon ein *grundsätzliches Problem* von Datenerhebungen erkennen. Wenn die Grundgesamtheit der erfaßten Personen bekannt ist und keine weiteren SDD erfaßt werden (dies entspricht ZZ = 1), so läßt sich dennoch jede Person mit der Wahrscheinlichkeit von 1/N re-individualisieren. Diese Wahrscheinlichkeit ist größer als Null (für $N \geqslant 1$) und muß noch zu dem prozentualen Anteil der Deanonymisierbarkeit auf Grund der einfach besetzten Zellen zugerechnet werden. Ist ZZ = 1 und $N > 1$, so ist EZ = 0 und es bleibt die Wahrscheinlichkeit von 1/N übrig.

Dieses Ergebnis besagt, daß ein *absoluter Schutz* nur dann zu gewährleisten ist, wenn überhaupt keine Daten erhoben werden oder der erfaßte Personenkreis unbekannt ist!

Von daher stellt sich nun das Problem, welches Ausmaß an Deanonymisierbarkeit ist noch zu tolerieren.

Wie groß DA sein wird, hängt von der Größe der erfaßten Population, der insgesamt bereitgestellten Zellenzahl ZZ und den Wahrscheinlichkeitsverteilungen über den einzelnen SDD ab.

Im weiteren werde ich insgesamt vier verschiedene Möglichkeiten zur Abschätzung von EZ vorstellen, diese an verschiedenen Zellenbesetzungen validieren, um sie zum Schluß den aufgestellten Bewertungskriterien zu unterziehen.

Aufstellung von Bewertungskriterien

Um die im nächsten Abschnitt vorgestellten Abschätzungsverfahren später sinnvoll bewerten zu können, werde ich drei Kriterien aufstellen:

1. *„Verankerung"*

   Für den Fall von ZZ = 1 soll das Abschätzverfahren die Rest-Wahrscheinlichkeit von 1/N liefern.

2. *„Natürliches Verhalten"*

   Hiermit ist gemeint, daß die einzelnen Werte für DA bei konstanter Verteilung mit wachsendem N möglichst genau den Werteverlauf im ungünstigsten Fall (d. h. das jeweilige Maximum für EZ) in seiner qualitativen Form wiedergeben. Dies läßt sich wie folgt formal darstellen:

   $$|DA_j\ (DAV) - DA_j\ (BU)| \approx \text{constant} \qquad (4)$$

   mit j = 1 ... N und für alle möglichen Verteilungen;
   DAV = *D*eanonymisierungs*a*bschätzungs*v*erfahren,
   BU = *b*eobachtete Zellenbesetzung im *u*ngünstigsten Fall.

3. *„Sicherheit"*

   Unter diesem Bewertungskriterium soll der Tatsache Rechnung getragen werden, daß die Abschätzung von DA möglichst den ungünstigsten Fall mit erfaßt. Dies stellt sich formal so dar:

   $$DA_j\ (DAV) \geqslant DA_j\ (BU) \qquad (5)$$

   mit j = 1 ... N und für alle möglichen Verteilungen.

Herleitung von vier Verfahren zur Abschätzung der Deanonymisierbarkeit (DAV)

Zur Beantwortung der aufgeworfenen Frage nach dem zu erwartenden Deanonymisierbarkeits-Risiko bei einer geplanten Datenerhebung stelle ich vier DAV vor, die dann an Hand eines Beispiels auf ihre Brauchbarkeit überprüft werden.

1. *DAV „Freiheitsgrade"*

   Dieses DAV (FG) ist das einfachste der vorgeschlagenen Abschätzungsverfahren und geht von dem Gedanken aus, daß die maximale Anzahl einfach besetzter Zellen von der Anzahl der im m-dimensionalen Datenkörper unter jeder beliebigen Wahrscheinlichkeitsverteilung zur Verfügung stehenden Freiheitsgraden für die Zellenbesetzungen abhängt.

   $$EZ\ (FG) = (n_m - 1)(n_{m-1} - 1) \prod_{i=1}^{m-2} n_i \text{ mit } n_i \leqslant n_{i+1} \qquad (6)$$

2. *DAV „Randverteilung – trunc"*

   Dieses DAV (RT) geht von der Überlegung aus, daß sich die Anzahl der einfach besetzten Zellen über die erwarteten Häufigkeiten (EH) auf Grund der Randverteilungen abschätzen läßt. Die Randverteilung der i-ten Dimension entspricht der Wahrscheinlichkeitsverteilung der i-ten SDD in der Grundgesamtheit. Im m-dimensionalen Fall sieht die Abschätzung der EH wie folgt aus:

   $$EH_{ijk\ldots l} = p_{1_i} * p_{2_j} * p_{3_k} * \ldots * p_{m_l} \qquad (7)$$

   mit $i = 1 \ldots n_1$
   $j = 1 \ldots n_2$
   $k = 1 \ldots n_3$
   $\vdots$
   $l = 1 \ldots n_m$

   Das Produkt aus $EH_{ijk\ldots l}$ mit N ergibt dann die absolute Häufigkeit der Zelle (ijk ... l). Da jedoch diese Berechnung keine ganzzahligen Werte für EH liefert, müssen sie noch einer weiteren Transformation unterzogen werden. Für dieses DAV habe ich die Transformation des rechts halboffenen Intervalls

   $$\text{trunc} : [x\ ; x + 1[\ \rightarrow x \qquad (8)$$

   gewählt. Damit lassen sich nun die einfach besetzten Zellen wie folgt abschätzen:

   $$EZ\ (RT) = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} \sum_{l=1}^{n_m} I * \text{trunc}\ (EH_{ijk\ldots l} * N) \qquad (9)$$

   Der Paramter I stellt lediglich eine Indikatorfunktion für die Summation dar, die folgende Bedingungen genügt:

$$I = \begin{cases} 1 \; ; \text{für trunc}\,(EH_{ijk\ldots l} * N) = 1 \\ 0 \; ; \text{sonst} \end{cases} \qquad (10)$$

Diese Indikatorfunktion trägt dafür Sorge, daß nur die nach DAV (RT) erwarteten *einfach* besetzten Zellen aufsummiert werden.

3. *DAV „Randverteilung – round"*

Ähnlich wie bei dem vorherigen DAV (RT) wird auch bei diesem DAV (RR) die Anzahl der EZ über das Produkt der einzelnen relativen Häufigkeiten der Randverteilungen abgeschätzt (siehe Formel 9).

Statt jedoch der Transformation 'trunc' wird bei diesem DAV (RR) die folgende Abbildung verwendet:

$$\text{round} : [x - 0.5 \; ; \; x + 0.5[\; \rightarrow x \qquad (11)$$

Es wird also die Anzahl der EZ wie folgt bestimmt:

$$EZ\,(RR) = \sum_{i=1}^{n_n} \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} \ldots \sum_{l=1}^{n_m} I * \text{round}\,(EH_{ijk\ldots l} * N) \qquad (12)$$

Für die Indikatorfunktion I gelten die gleichen Bedingungen wie in Formel 10, wobei jedoch 'trunc' durch 'round' ersetzt werden muß.

4. *DAV „ORWELL-Koeffizient"*

Für dieses Abschätzungsverfahren DAV (OR) wird zunächst von der mittleren Ungewißheit U_i der Wahrscheinlichkeitsverteilung der i-ten SDD ausgegangen, die sich wie folgt bestimmen läßt:

$$U_i = -\sum_{j=1}^{n_i} p_{i_j} * \log_B p_{i_j} \quad \text{mit } B = 2 \text{ als Basis und } i = 1 \ldots m \qquad (13)$$

Um nun daraus ein Maß für die im Mittel real genutzten Ausprägungsstufen (A_i) der i-ten SDD zu berechnen, werden zunächst folgende Anforderungen an dieses Maß aufgestellt:

a) für $p_{i_j} = 1.0$ ergibt sich $A_i = 1; j = n_i = 1$
b) für $p_{i_j} = 1/n_i$ ergibt sich $A_i = n_i; j = 1 \ldots n_i$
c) A_i steht in einem monoton wachsendem Verhältnis zu U_i, wenn n_i konstant bleibt;
d) A_i steht in einem monoton wachsenden Verhältnis zu n_i, wenn U_i konstant bleibt.

Die folgende Formel 14 erfüllt diese Anforderungen:

$$A_i = B^{U_i} \quad \text{für } i = 1 \ldots m \qquad (14)$$

Um zu einer einfach zu handhabenden Berechnungsvorschrift für die einzelnen A_i zu gelangen, ergibt sich nach Einsetzen und entsprechender Umformung die folgende Berechnungsvorschrift:

$$A_i = \left[\prod_{j=1}^{n_i} p_{i_j}^{\;p_{i_j}}\right]^{-1} \quad \text{mit } i = 1 \ldots m \qquad (15)$$

Dieses Maß für die im Mittel real genutzten Ausprägungsstufen der i-ten SDD erfüllt nun folgende Bedingung:

$$1 \leqslant A_i \leqslant n_i \quad \text{für alle } i = 1 \ldots m \qquad (16)$$

Die Gleichheit ($A_i = n_i$) gilt dann, wenn die einzelnen Ausprägungsstufen der i-ten SDD in der Grundgesamtheit gleichverteilt sind.

Die Anzahl der einfach besetzten Zellen läßt sich nun wie folgt abschätzen:

$$EZ\,(OR) = \prod_{i=1}^{m} A_i \quad \text{„ORWELL-Koeffizient"} \qquad (17)$$

Damit wären wir nun soweit, die Brauchbarkeit der vier DAV zu überprüfen.

Überprüfung der Abschätzungsverfahren auf ihre Brauchbarkeit

An einem einfachen Beispiel werde ich die Ergebnisse der Abschätzung von EZ darstellen.

Gehen wir von dem nicht unrealistischen Fall aus, daß in einer Behörde beschlossen wird, über einen bestimmten Bereich der Realität Daten zu Planungszwecken zu erheben. Da jedoch der zu erhebende Datensatz später als *nicht eindeutig* deanonymisierbar gelten soll, stellt sich zunächst die Frage, wie genau die einzelnen SDD zu erheben sind, so daß diese in ihrer Gesamtheit bei einem vorgegebenen N und den bekannten Verteilungen der einzelnen SDD in der Grundgesamtheit einem aus politischen und/oder juristischen Überlegungen heraus festzulegenden Deanonymisierbarkeitsgrad (DA) genüge tun.

Zunächst ist die Erhebung im soziodemographischen Bereich mit den drei Fragen geplant:

Alter
mit den vier Ausprägungsstufen
(< 20) (20–40) (41–60) (> 60);

Schulbildung
mit den drei Ausprägungsstufen
(Volksschule) (Mittelschule) (Oberschule);

Geschlecht
mit den zwei Ausprägungsstufen
(weiblich) (männlich)

Für ein gegebenes N und der erst einmal angenommenen Gleichverteilung für jede dieser drei SDD habe ich nun die absoluten Häufigkeiten (H) bestimmt und in ein Schema nach Abbildung 1 eingetragen.

Um nun einen etwas allgemeineren Überblick über die Ergebnisse der einzelnen DAV zu bekommen, habe ich dies für eine Reihe anderer Wahrscheinlichkeitsverteilungen wiederholt (siehe Tabelle 1 bis 9). Zusätzlich habe ich drei Beispiele schematisch (Abbildung 1, 3 und 5) und graphisch (Abbildung 2, 4 und 6) zur Veranschaulichung dargestellt.

Für die neun verschiedenen, vorgegebenen Wahrscheinlichkeitsverteilungen lassen sich nun die einzelnen Abschätzungen der vier DAV mit der ‚realen', beobachtbaren Zellenbesetzung für verschiedene Populationsgrößen N vergleichen.

Betrachtet man nun die Tabellen 1 bis 8 und die drei exemplarisch ausgewählten Graphiken, so zeigt sich, daß der ORWELL-Koeffizient deutlich dem Kriterium der „Sicherheit" genüge tut. Bis auf eine Ausnahme in Tabelle 6 zeigt der ORWELL-Koeffizient eine eindeutige Überlegenheit den drei anderen Abschätzungsverfahren gegenüber.

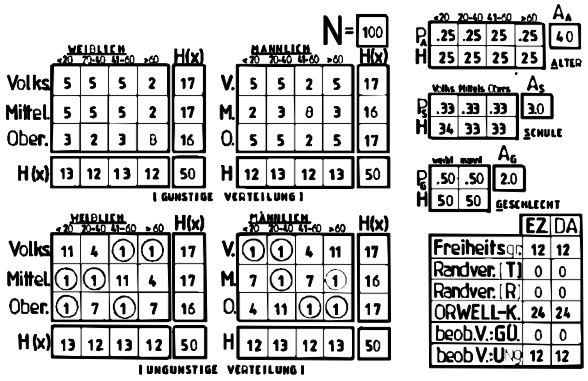


Abb. 1

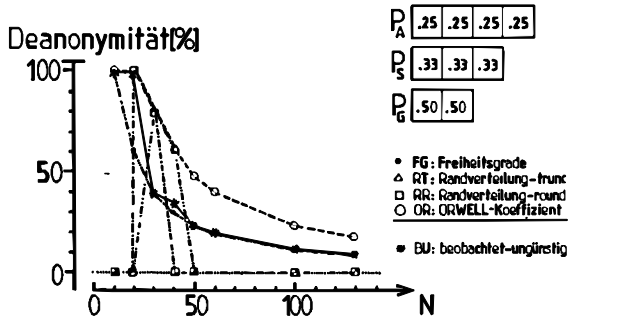


Abb. 2 Funktionale Abhängigkeit der einzelnen Deanonymisierbarkeits-Abschätzungsverfahren von der Größe der Population (N) im Vergleich mit den realen Verteilungen

Tabelle 1 Ergebnisse der Deanonymisierbarkeit-Abschätzungen nach den vier DAV unter gegebener Wahrscheinlichkeitsverteilung für acht verschiedene Populationsgrößen (N) im Vergleich zu der realen Zellenbesetzung

P_{i_j}	Verteilung	N	FG		RT		RR		OR		Beobachtet günstig		Beobachtet ungünstig	
			EZ	DA	EZ	DA	EZ	DA	EZ	DA	EZ	DA	EZ	DA
P_{G_j}	.50 .50	10	12	100	0	0	0	0	24	100	6	60	10	100
		20	12	60	0	0	24	100	24	100	6	30	20	100
P_{S_j}	.33 .33 .33	30	12	40	24	80	24	80	24	80	2	5	16	40
		40	12	30	24	60	0	0	24	60	0	0	14	35
P_{A_j}	.25 .25 .25 .25	50	12	24	0	0	0	0	24	48	0	0	12	24
		60	12	20	0	0	0	0	24	40	0	0	12	20
		100	12	12	0	0	0	0	24	24	0	0	12	12
		130	12	9	0	0	0	0	24	18	0	0	12	9

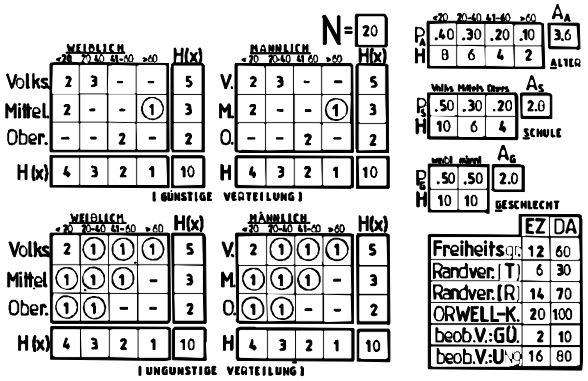


Abb. 3

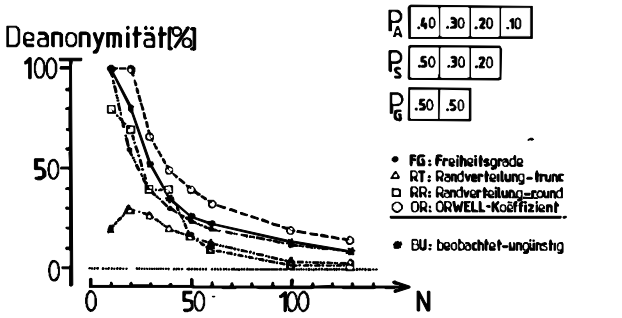


Abb. 4 Funktionale Abhängigkeit der einzelnen Deanonymisierbarkeits-Abschätzungsverfahren von der Größe der Population (N) im Vergleich mit den realen Verteilungen

Tabelle 2 Ergebnisse der Deanonymisierbarkeits-Abschätzungen nach den vier DAV unter gegebener Wahrscheinlichkeitsverteilung für acht verschiedene Populationsgrößen (N) im Vergleich zu der realen Zellenbesetzung

P_{i_j}	Verteilung	N	FG		RT		RR		OR		Beobachtet günstig		Beobachtet ungünstig	
			EZ	DA	EZ	DA	EZ	DA	EZ	DA	EZ	DA	EZ	DA
P_{G_j}	.50 .50	10	12	100	0	0	6	60	22	100	4	40	10	100
		20	12	60	6	30	18	90	22	100	2	10	16	80
P_{S_j}	.33 .33 .33	30	12	40	12	40	12	40	22	73	3	10	13	43
		40	12	30	12	30	12	30	22	55	0	0	14	35
P_{A_j}	.40 .30 .20 .10	50	12	24	6	12	6	12	22	44	0	0	14	28
		60	12	20	6	10	6	10	22	37	0	0	14	23
		100	12	12	6	6	0	0	22	22	0	0	12	12
		130	12	9	0	0	0	0	22	17	0	0	12	9

Tabelle 3 Ergebnisse der Deanonymisierbarkeits-Abschätzungen nach den vier DAV unter gegebener Wahrscheinlichkeitsverteilung für acht verschiedene Populationsgrößen (N) im Vergleich zu der realen Zellenbesetzung

P_{i_j}	Verteilung	N	FG		RT		RR		OR		Beobachtet günstig		Beobachtet ungünstig	
			EZ	DA	EZ	DA	EZ	DA	EZ	DA	EZ	DA	EZ	DA
P_{G_j}	.50 .50	10	12	100	2	20	8	80	20	100	6	60	10	100
		20	12	60	6	30	14	70	20	100	2	10	16	80
P_{S_j}	.50 .30 .20	30	12	40	8	27	12	40	20	67	1	3	16	53
		40	12	30	8	20	10	40	20	50	0	0	14	35
P_{A_j}	.40 .30 .20 .10	50	12	24	8	16	8	16	20	40	0	0	13	26
		60	12	20	8	13	6	10	20	33	0	0	14	23
		100	12	12	4	4	2	2	20	20	0	0	14	14
		130	12	9	4	3	2	2	20	15	0	0	12	9

Tabelle 4 Ergebnisse der Deanonymisierbarkeit-Abschätzungen nach den vier DAV unter gegebener Wahrscheinlichkeitsverteilung für acht verschiedene Populationsgrößen (N) im Vergleich zu der realen Zellenbesetzung

P_{i_j}	Verteilung	N	FG		RT		RR		OR		Beobachtet günstig		Beobachtet ungünstig	
			EZ	DA	EZ	DA	EZ	DA	EZ	DA	EZ	DA	EZ	DA
P_{G_j}	.50 .50	10	8	80	2	20	6	60	14	100	5	50	8	80
		20	8	40	6	30	8	40	14	70	4	20	12	60
P_{S_j}	.50 .30 .20	30	8	27	4	13	6	20	14	47	1	3	9	30
		40	8	20	6	15	6	15	14	35	0	0	8	20
P_{A_j}	.60 .30 .10 –	50	8	16	4	8	6	12	14	28	0	0	9	18
		60	8	13	4	7	4	7	14	23	0	0	10	17
		100	8	8	4	4	2	2	14	14	0	0	8	8
		130	8	6	4	3	2	2	14	11	0	0	8	6

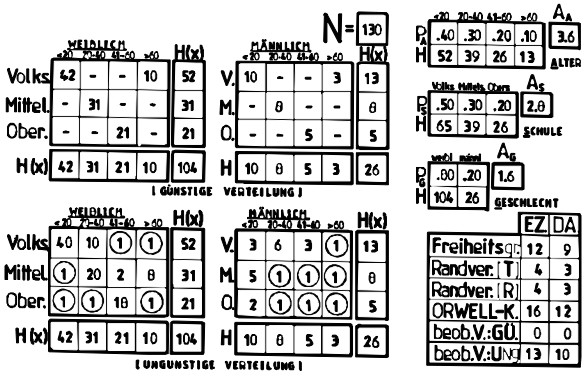


Abb. 5

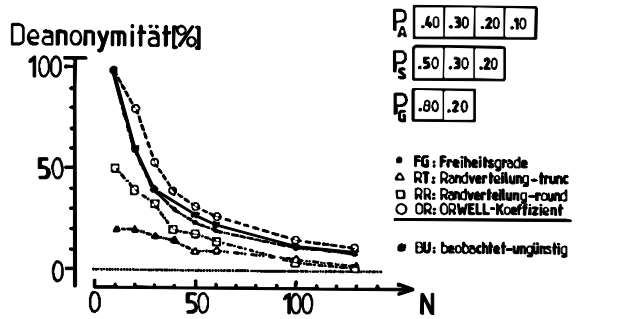


Abb. 6 Funktionale Abhängigkeit der einzelnen Deanonymisierbarkeits-Abschätzungsverfahren von der Größe der Population (N) im Vergleich mit den realen Verteilungen

Tabelle 5 Ergebnisse der Deanonymisierbarkeits-Abschätzungen nach den vier DAV unter gegebener Wahrscheinlichkeitsverteilung für acht verschiedene Populationsgrößen (N) im Vergleich zu der realen Zellenbesetzung

P_{i_j}	Verteilung	N	FG		RT		RR		OR		Beobachtet günstig		Beobachtet ungünstig	
			EZ	DA	EZ	DA	EZ	DA	EZ	DA	EZ	DA	EZ	DA
P_{G_j}	.80 .20	10	12	100	2	20	5	50	16	100	5	50	10	100
		20	12	60	4	20	8	40	16	80	2	10	12	60
P_{S_j}	.50 .30 .20	30	12	40	5	17	10	33	16	53	2	7	12	40
		40	12	30	6	15	8	20	16	40	1	3	15	38
P_{A_j}	.40 .30 .20 .10	50	12	24	5	10	9	18	16	32	1	2	14	28
		60	12	20	6	10	9	15	16	27	0	0	14	23
		100	12	12	6	6	5	5	16	16	0	0	13	13
		130	12	9	4	3	4	3	16	12	0	0	13	10

Tabelle 6 Ergebnisse der Deanonymisierbarkeits-Abschätzungen nach den vier DAV unter gegebener Wahrscheinlichkeitsverteilung für acht verschiedene Populationsgrößen (N) im Vergleich zu der realen Zellenbesetzung

P_{I_j}	Verteilung	N	FG		RT		RR		OR		Beobachtet günstig		Beobachtet ungünstig	
			EZ	DA	EZ	DA	EZ	DA	EZ	DA	EZ	DA	EZ	DA
P_{G_j}	.80 .20	10	8	80	2	20	4	40	10	100	3	30	7	70
		20	8	40	2	10	6	30	10	50	3	15	7	35
P_{S_j}	.60 .30 .10	30	8	27	4	13	7	23	10	33	0	0	9	30
		40	8	20	4	10	3	8	10	25	1	3	11	28
P_{A_j}	.60 .30 .10 –	50	8	16	4	8	5	10	10	20	1	2	9	18
		60	8	13	3	5	5	8	10	17	1	2	7	12
		100	8	8	3	3	5	5	10	10	0	0	9	9
		130	8	6	3	2	3	2	10	8	0	0	8	6

Tabelle 7 Ergebnisse der Deanonymisierbarkeits-Abschätzungen nach den vier DAV unter gegebener Wahrscheinlichkeitsverteilung für acht verschiedene Populationsgrößen (N) im Vergleich zu der realen Zellenbesetzung

P_{I_j}	Verteilung	N	FG		RT		RR		OR		Beobachtet günstig		Beobachtet ungünstig	
			EZ	DA	EZ	DA	EZ	DA	EZ	DA	EZ	DA	EZ	DA
P_{G_j}	1.0 –	10	6	60	3	30	5	50	9	90	1	10	7	70
		20	6	30	3	15	5	25	9	45	0	0	6	30
P_{S_j}	.60 .30 .10	30	6	20	3	10	4	13	9	30	0	0	7	23
		40	6	15	3	8	3	8	9	23	0	0	7	18
P_{A_j}	.40 .30 .20 .10	50	6	12	3	6	2	4	9	18	0	0	6	12
		60	6	10	3	5	2	4	9	15	0	0	6	10
		100	6	6	1	1	1	1	9	9	0	0	6	6
		130	6	5	1	1	1	1	9	7	0	0	6	5

Tabelle 8 Ergebnisse der Deanonymisierbarkeits-Abschätzungen nach den vier DAV unter gegebener Wahrscheinlichkeitsverteilung für acht verschiedene Populationsgrößen (N) im Vergleich zu der realen Zellenbesetzung

P_{I_j}	Verteilung	N	FG		RT		RR		OR		Beobachtet günstig		Beobachtet ungünstig	
			EZ	DA	EZ	DA	EZ	DA	EZ	DA	EZ	DA	EZ	DA
P_{G_j}	1.0 –	10	4	40	2	20	3	30	6	60	1	10	4	40
		20	4	20	3	15	4	20	6	30	0	0	5	25
P_{S_j}	.60 .30 .10	30	4	13	2	7	2	7	6	20	0	0	6	20
		40	4	10	2	5	2	5	6	15	0	0	4	10
P_{A_j}	.60 .30 .10 –	50	4	8	2	4	1	2	6	12	0	0	4	8
		60	4	7	2	3	1	2	6	10	0	0	4	7
		100	4	4	1	1	1	1	6	6	0	0	4	4
		130	4	3	1	1	1	1	6	5	0	0	4	3

Tabelle 9 Ergebnisse der Deanonymisierbarkeits-Abschätzungen nach den vier DAV unter gegebener Wahrscheinlichkeitsverteilung für acht verschiedene Populationsgrößen (N) im Vergleich zu der realen Zellenbesetzung

P_{I_j}	Verteilung	N	FG		RT		RR		OR		Beobachtet günstig		Beobachtet ungünstig	
			EZ	DA	EZ	DA	EZ	DA	EZ	DA	EZ	DA	EZ	DA
P_{G_j}	1.0 –	10	0	0	0	0	0	0	1	10	0	0	0	0
		20	0	0	0	0	0	0	1	5	0	0	0	0
P_{S_j}	1.0 – –	30	0	0	0	0	0	0	1	3	0	0	0	0
		40	0	0	0	0	0	0	1	3	0	0	0	0
P_{A_j}	1.0 – – –	50	0	0	0	0	0	0	1	2	0	0	0	0
		60	0	0	0	0	0	0	1	2	0	0	0	0
		100	0	0	0	0	0	0	1	1	0	0	0	0
		130	0	0	0	0	0	0	1	1	0	0	0	0

Ebenso wie DAV (FG) zeigt der ORWELL-Koeffizient ein weitgehend „natürliches Verhalten".

Die überhöhte Abschätzung der Deanonymisierbarkeit durch den ORWELL-Koeffizienten läßt sich unter Einbeziehung der Ergebnisse aus Tabelle 9 und dem oben angesprochenen Grundproblem auf das in dieser Abhandlung gewählte Kriterium der Deanonymisierbarkeit (die Bestimmung der Anzahl an einfach besetzten Zellen) zurückführen. In Wirklichkeit muß über die Berücksichtigung der einfach besetzten Zellen hinaus, die in der Tat eine *eindeutige* Deanonymisierung möglich machen, auch der Tatsache Rechnung getragen werden, daß bei der zufälligen Auswahl von einer Person aus den N Personen die Erfolgswahrscheinlichkeit einer zutreffenden Auswahl 1/N beträgt.

Diese Tatsache wurde bei der Konstruktion des ORWELL-Koeffizienten durch die Bedingung a) berücksichtigt. Somit erfüllt der ORWELL-Koeffizient als einziges der vorgestellten DAV das Kriterium der „Verankerung".

Insgesamt läßt sich somit feststellen, daß das Abschätzungsverfahren nach dem ORWELL-Koeffizienten das zur Zeit beste Ergebnis auf die eingangs aufgeworfene Frage liefert.

Zum Schluß sei noch einmal darauf hingewiesen, daß es hierbei um die Abschätzung von Datenerhebungen im *vorhinein* geht und nicht um die Abschätzung der Deanonymisierbarkeit von schon erhobenen Datenbeständen. Es wird hier davon ausgegangen, daß die geplante Erfassung an soziodemographischen Daten in der speziellen ausgewählten Zusammenstellung bis dahin noch nicht vorliegt oder aus irgendwelchen Gründen nicht zugänglich ist.

Bei schon erhobenen Datenbeständen stellt sich das Problem der Bestimmung der Anzahl an einfach besetzten Zellen in weitgehend trivialer Weise durch Auszählen! Sollte das Auszählen aus irgendwelchen Gründen jedoch nicht möglich sein, so ist auch in diesem Falle die Abschätzung der Deanonymisierbarkeit unter zu Hilfenahme des ORWELL-Koeffizienten angezeigt.

Der Vorteil des ORWELL-Koeffizienten liegt also primär darin, bei der Planung und Ausgestaltung von Erhebungsbögen dem Datenerheber — unter einer vorgegebenen Deanonymisierungsrate — eine Richtlinie für die Bestimmung der Anzahl an Ausprägungsstufen der einzelnen soziodemographischen Fragen an die Hand zu geben.

Anwendung des ORWELL-Koeffizienten

In der Planungsphase einer größeren Datenerhebung werden zunächst einmal die in Frage kommenden soziodemographischen Fragen zusammengestellt und die jeweiligen Ausprägungsstufen festgelegt.

Sodann werden die dazugehörigen relativen Häufigkeiten aus den bekannten Statistiken entnommen und für jede der soziodemographischen Fragen unter der gegebenen Verteilung nach Formel 15 die im Mittel real genutzten Ausprägungsstufen errechnet. Falls keine relativen Häufigkeiten für die eine oder andere soziodemographische Frage zugänglich und keine plausiblen Annahmen möglich sind, so sollte man die Gleichverteilung als die konservative Schätzung den Berechnungen zu Grunde legen.

Danach erhält man über Formel 17 („ORWELL-Koeffizient") und Formel 3 den prozentualen Anteil an deanonymisierbare Personen.

Wenn dieser Prozentsatz zu hoch ausfallen sollte, so verkleinert man die Anzahl der Ausprägungsstufen und berechnet unter Beachtung der veränderten relativen Häufigkeiten den ORWELL-Koeffizienten erneut. (Das Problem kann aber auch ebenso über die Erhöhung der Anzahl an erfaßten Personen gelöst werden.)

Dies Vorgehen wiederholt man solange, bis die erwünschte Deanonymisierungs-Rate unterschritten wurde.

Die zu fordernde „ausreichende" Anonymisierung von Datensätzen

Hans Hammerbacher, Wiesbaden

Ging der vorhergehende Beitrag von der Prämisse aus, daß jedem das Zusatzwissen zur Individualisierung anonymisierter Daten zur Verfügung steht und er daran auch ein absolutes Interesse hat, geht dieser Beitrag von dem Umstand aus, daß „personenbezogen" und „anonymisiert" einander nicht ausschließen, sondern ohne deutliche objektive Grenze ineinander übergehen. Die Grenze soll nach Ansicht des Verfassers durch das dem Betroffenen schädliche Interesse an den Daten bestimmt werden; die Anonymisierung soll diesem schädlichen Interesse gezielt vorbeugen.

Im Gegensatz zum ersten Beitrag wird hier verlangt, daß nicht die Datei pauschal zu bewerten, sondern daß jeder einzelne zu einer Person gespeicherte Datensatz zu prüfen ist.

Die Herausgeber