3 Methodische Grundlagen von Reidentifikationsversuchen*

Nachdem im vorangegangenen Kapitel verschiedene Forschungsschwerpunkte zu Deanonymisierungsrisiken dargestellt wurden, sollen im folgenden die Grundprinzipien einer Reidentifikation sowie die hierbei entstehenden Probleme präzisiert werden. Ein wesentlicher Teil dieses Kapitels dient der Darstellung unterschiedlicher Reidentifikationstechniken und der Einschätzung ihrer Leistungsfähigkeit. Hierbei wird zum einen die Methode des einfachen Abgleichs (Abschnitt 3.2), zum anderen die bereits mehrfach erwähnte diskriminanzanalytische Reidentifikationstechnik von Paaß/Wauschkuhn (Abschnitt 3.4) detailliert dargestellt, da ihre jeweilige Leistungsfähigkeit im weiteren empirisch überprüft werden soll. Darüber hinaus soll mit dem sogenannten Distanzminimierungsverfahren eine weitere mögliche Reidentifikationstechnik (Abschnitt 3.3) erörtert werden, die jedoch als "Zwischenstück" von einfacher Abgleichtechnik und diskriminanzanalytischer Reidentifikationstechnik keiner gesonderten empirischen Überprüfung unterzogen werden soll.

3.1 Grundprinzipien einer Reidentifikation

Das Grundprinzip einer Reidentifikation beruht darauf, daß Einzeldatensätze einer Datei mit anonymen Mikrodaten (Mikrodatenfile) den Einzeldatensätzen einer anderen personenbezogenen Datei in einer eins-zu-eins Entsprechung zugeordnet werden (vgl. u.a. Schlörer 1980, Dittrich/Schlörer 1987). Die für einen Reidentifikationsversuch unabdingbare personenbezogene Datei wird als Identifikationsfile bezeichnet. Sie entspricht dem sogenannten Zusatzwissen (vgl. Kapitel 6) und muß sich - zumindest in Teilen - auf die gleichen Personen wie das anonymisierte Mikrodatenfile beziehen. Im Extremfall kann dieses Zusatzwissen aus einem einzigen Datensatz bestehen, der nicht maschinenlesbar dargestellt sein muß.

Deanonymisierungsversuche setzen bei den Merkmalen an, die Zusatzwissen und Mikrodatenfile gemeinsam sind (Überschneidungsmerkmale). Durch einen Vergleich der Merkmalsausprägungen der Einzeldatensätze beider Files wird

---

* Autoren: Heike Wirth (3.1, 3.2)
Uwe Blien, Heike Wirth (3.3)
Stefan Bender, Uwe Blien (3.4)

hierbei angestrebt, auf der Basis identischer oder sehr ähnlicher Werte jene Datensätze zuzuordnen, die von ein und derselben Person stammen. Damit können die im Mikrodatenfile enthaltenen Nutzmerkmale (beispielsweise Einkommen, Schulden etc.) identifizierbaren Personen zugewiesen werden. Ein Ziel eines solchen Reidentifikationsversuchs könnte darin bestehen, das daraus resultierende personenbezogene Wissen für wissenschaftliche oder andere Zwecke zu nutzen (vgl. Übersicht 3.1).

Übersicht 3.1: Die einem Reidentifikationsversuch zugrundeliegende Datenbasis

	IDENTIFIKATIONS-MERKMALE (z.B. Name, Anschrift)	ÜBERSCHNEIDUNGS-MERKMALE (z.B. Beruf, Alter)	NUTZMERKMALE (z.B. Einkommen)
MIKRODATENFILE	fehlen	vorhanden	vorhanden
IDENTIFIKATIONSFILE (ZUSATZWISSEN)	vorhanden	vorhanden	fehlen

Bei Deanonymisierungsversuchen können im wesentlichen zwei Strategien[2] angewandt werden: Im Falle einer gezielten Suche (*Einzelsuche*) sollen bestimmte, dem Angreifer bekannte Personen aus einem Mikrodatenfile deanonymisiert werden. Bei einem *Fischzug* dagegen werden nicht vorgegebene, sondern beliebige Personen in einem Mikrodatenfile gesucht (Paaß/Wauschkuhn 1985).

Ein Reidentifikationsversuch ist dann relativ problemlos, wenn beide benötigten Datenfiles, also Zusatzwissen und Mikrodatenfile, die in ihnen enthaltenen Einzelfälle in den Datensätzen völlig kompatibel abbilden und mindestens einer der beiden Datenfiles die gesamte Population enthält. In diesem Fall muß lediglich geprüft werden, ob sich für einen Einzelfall des Identifikationsfile ein in den Merkmalsausprägungen identischer Fall im Mikrodatenfile findet und die Beziehung eindeutig ist. Ist eine solche eins-zu-eins Zuordnung (im folgenden als *eindeutige Zuordnung* bezeichnet) möglich, kann der Angreifer zweifelsfrei sicher sein, daß die zugeordneten Datensätze von ein und derselben Person stammen. Da für den Datensatz des Mikrodatenfile damit ein eindeutiger Per-

2 Die sich aus den unterschiedlichen Strategien für die Reidentifikationswahrscheinlichkeit ergebenden Konsequenzen werden in Kapitel 4 ausführlich erläutert.

sonenbezug hergestellt wäre, würde eine Reidentifikation im Sinne des BStatG vorliegen.

Unter den gleichen Annahmen (keine Inkompatibilitäten, Mikrodatenfile und/oder Identifikationsfile ist eine Vollerhebung) kann ein eindeutiger Personenbezug allerdings schon dann nicht mehr hergestellt werden, wenn eine spezifische Ausprägungskombination im Mikrodatenfile und/oder Identifikationsfile mehrfach besetzt ist, d.h. *statistische Doppelgänger* auftreten. Liegt aufgrund von statistischen Doppelgängern eine Zuordnungsrelation von 1:n, n:1 oder n:n, d.h. eine *mehrdeutige Zuordnung* vor, ist ein Angreifer ohne zusätzliche Informationen nicht in der Lage, für einen Einzeldatensatz einen spezifischen Personenbezug herzustellen. Zusatzwissen steht in aller Regel jedoch nicht unbegrenzt zur Verfügung. Wäre dies der Fall würde sich ein Reidentifikationsversuch erübrigen. Aus diesem Grund werden mehrdeutige Zuordnungen im allgemeinen als weniger riskant und daher als untergeordnetes Problem aus der Diskussion ausgeklammert (Paaß/Wauschkuhn 1985).

Vorwiegend in der englischsprachigen und vereinzelt auch in der deutschsprachigen Literatur wird der Deanonymisierungsbegriff allerdings zum Teil in einem umfassenderen Sinn einer Aufdeckung bzw. Enthüllung (*Disclosure*) von Informationen verwandt, bei welcher nicht notwendigerweise eine Reidentifikation (d.h. die Herstellung eines Personenbezugs für einen spezifischen Datensatz) vorliegen muß (Dalenius 1977, Schlörer 1980, eine zusammenfassende Darstellung findet sich bei Duncan/Lambert 1987). Die oben gegebene Definition der eindeutigen Zuordnung zweier Datensätze aus unterschiedlichen Datenfiles mit dem Ziel einer Reidentifikation einer oder mehrerer spezifischen/spefizischer Person(en) wird in diesen Konzept[3] als sogenannte 'identity disclosure' (vereinzelt auch als 'personal disclosure') bezeichnet (vgl. unter anderem Bethlehem et al. 1990, Paaß 1988a, Skinner et al. 1990, Marsh et al. 1991). Von 'statistical disclosure' wird dann gesprochen, wenn es gelingt einen Einzeldatensatz korrekt mit anderen zuverlässigen Informationen zu verknüpfen, d.h. also eine Datensatzerweiterung vorzunehmen, auch wenn die hinter diesem Einzeldatensatz stehende Person nicht bekannt ist (Cox/Sande 1979). Dalenius (1977) spricht von 'attribute disclosure', wenn es durch die in

3 Wenn hier von einem Konzept gesprochen wird, dann in dem Sinne, daß unterschiedlichste Formen von Informationsenthüllung unter dem Oberbegriff Disclosure zusammengefaßt werden. Sowohl Schlörer (1980) wie auch Duncan/Lambert (1987) weisen darauf hin, daß unterschiedliche Autoren ein und denselben Begriff für unterschiedliche Sachverhalte verwenden und daher keinesfalls ein in sich geschlossenes Disclosure Konzept zur Verfügung steht.

einem Mikrodatenfile enthaltenen Informationen möglich ist, Merkmalsausprägungen für einen spezifischen Datensatz einer Person genauer zu bestimmen, als dies ohne die in dem Mikrodatenfile enthaltenen Informationen möglich wäre. Auch hier ist die Identität der Person ohne Belang. Palley und Simonoff (1986) schließlich beziehen 'disclosure' nicht mehr auf einzelne Fälle sondern auf spezifische Subpopulationen ('population disclosure'). 'Population disclosure' würde dann vorliegen, wenn es gelingt eine Verbindung zwischen spezifischen Gruppencharakteristika (als Beispiel wird die Verbindung von Lohn- und Angestelltencharakteristika angeführt) herzustellen, und man auf diese Weise zu statistischen Aussagen über bestimmte Personengruppen gelangt.

Wie aus dieser äußerst knappen Darstellung unterschiedlicher disclosure-Typen hervorgeht, muß sich disclosure nicht auf erkennbare Einzelpersonen beziehen. Prinzipiell kann jede Form einer Enthüllung von statistischen Daten sowohl bezogen auf Gruppen mit mehreren tausenden Mitgliedern aber auch auf Einzelpersonen als 'disclosure' bezeichnet werden. Da der Zuordnungsbegriff im Sinne des BStatG die Herstellung eines eindeutigen Personenbezugs voraussetzt, ist daher nur die 'identity' bzw. 'personal disclosure', die der oben gegebenen Definition der eindeutigen Zuordnung entspricht, für die Bestimmung der faktischen Anonymität von Relevanz. In bezug auf eine Einordnung der mehrdeutigen Zuordnungen, muß allerdings auch die sogenannte 'attribute disclosure', d.h. die Enthüllung von Ausprägungen in aller Kürze erörtert werden.

Wie oben ausgeführt, bedeutet eine mehrdeutige Zuordnung, daß eine spezifische Ausprägungskombination in den Überschneidungsmerkmalen $Ü_1$ bis $Ü_n$ nicht nur auf eine Person, sondern auf n Personen zutrifft. Um einen spezifischen, eindeutigen Personenbezug im Sinne einer Reidentifikation herzustellen, müßte ein Angreifer über weitergehende Information verfügen. Dies trifft insbesondere dann zu, wenn die mehrdeutig zugeordneten Datensätze zwar für $Ü_1$ bis $Ü_n$ identische Ausprägungskombinationen aufweisen, aber sich in den Nutzmerkmalen N_1 bis N_n unterscheiden. Hiervon zu unterscheiden ist allerdings der Sonderfall, bei welchem ein Angreifer aus mehrdeutigen Zuordnungen auch ohne Herstellung eines eindeutigen Personenbezugs zumindest einen Informationsgewinn erhalten könnte.

Bei dieser spezifischen Bedingungskonstellation, die unter der erwähnten 'attribute disclosure' subsumiert werden könnte, ist vorausgesetzt, daß alle in einer mehrdeutigen Zuordnung enthaltenen, statistischen Doppelgänger, nicht

nur in den Überschneidungsmerkmalen $Ü_1$ bis $Ü_n$ sondern mindestens auch in einem der Nutzmerkmale N_1 bis N_n identisch sind (vgl. Übersicht 3.2).

Übersicht 3.2: Enthüllung von Ausprägungen bei mehrdeutigen Zuordnungen

IFz	$Ü_1$	$Ü_2$	...	$Ü_n$	MFy	N_1	N_2	...N_n
z_1	0	1	...	0	y_1	2	.	....
	0	1	...	0	y_2	2	.	....
	0	1	...	0	y_3	2	.	....
.					.			
.					.			
z_6	1	0	...	1	y_6	1	.	....
z_7	1	0	...	1	y_7	1	.	....
z_8	1	0	...	1	y_8	1	.	....

Unter dieser Bedingung könnte ein Angreifer für eine bestimmte Person die betreffende Ausprägung des Nutzmerkmals erschließen, auch wenn es ihm nicht möglich ist, für den spezifischen Datensatz einen Personenbezug herzustellen. Unter der Annahme, daß nicht nur *ein* Nutzmerkmal, sondern *alle* Nutzmerkmale N_1 bis N_n identisch wären, könnte ein Angreifer allerdings auch ohne weiteren Aufwand auf die Nutzmerkmale der ihn interessierenden Person schließen. Allerdings handelt es sich hierbei eher um eine Modellannahme, die für analytische Zwecke konstruiert ist, als um ein in der Realität häufig anzutreffendes Ereignis. Denn mit jedem zusätzlich einbezogenen Merkmal steigt die Wahrscheinlichkeit, daß sich zwei Datensätze bezüglich dieser Information unterscheiden. Daher ist es äußerst unwahrscheinlich, daß mehrdeutige zugeordnete Datensätze auch über alle Nutzmerkmale N_1 bis N_n identische Ausprägungen aufweisen. Eine bloße Enthüllung von Merkmalsausprägungen, die sich nicht - wie bei einer eindeutigen Zuordnung - erkennbar auf eine spezifische Person, sondern auf mehrere Personen beziehen, kann noch nicht als eine Reidentifikation im Sinne des BStatG bezeichnet werden, da diese immer die Herstellung eines eindeutigen Personenbezugs erfordert.

Vor diesem Hintergrund können sich die weitergehenden Erörterungen auf eindeutige Zuordnungen beschränken, wobei die in diesem Fall auftretenden Probleme analog auch bei mehrdeutigen Zuordnungen gegeben sind.

Walter Müller, Uwe Blien, Peter Knoche, Heike Wirth
unter Mitarbeit von
Petra Beckmann, Stefan Bender, Thomas Helmcke
und Michael Müller

Die faktische Anonymität von Mikrodaten

Band 19 der Schriftenreihe
Forum der Bundesstatistik
herausgegeben vom
Statistischen Bundesamt