11 Anonymisierungsmaßnahmen*

Für alle im vorausgegangenen Kapitel untersuchten Szenarien war das Kriterium des unverhältnismäßig hohen Aufwands mit Sicherheit erfüllt ,und damit die faktische Anonymität der Daten gegeben. Die empirische Überprüfung von Subszenario 1 und 2 auf der Basis von empirischen Daten konnte darüber hinaus aufzeigen, daß unter allen untersuchten Randbedingungen die Wahrscheinlichkeit einer eindeutigen und zugleich korrekten Zuordnung mit Werten zwischen 0 und 0,0005 empirisch wesentlich niedriger anzusetzen ist, als dies nach den - auf der Grundlage von synthetisch generiertem Zusatzwissen - ermittelten Befunden der GMD der Fall war. Sie haben auch gezeigt, daß aus einer hohen Quote von Fällen mit einzigartigen Ausprägungskombinationen noch keineswegs auf ein hohes Reidentifikationsrisiko geschlossen werden kann.

Eine wesentliche Erkenntnis war hierbei, daß das Ausmaß der von *Inkompatibilitäten* ausgehenden Schutzwirkung vor erfolgreichen Reidentifikationsversuchen bislang erheblich unterschätzt wurde (vgl. u.a. Bethlehem et al. 1990, Dalenius 1986, Brunnstein 1987). Während aufgrund der theoretischen Vorüberlegungen zu den Zuordnungsmechanismen einfacher Abgleichtechniken bereits deutlich wurde, daß diese Algorithmen beim Auftreten von Inkompatibilitäten mit "Nicht-" und/oder "Falschzuordnungen" reagieren, überraschen die Ergebnisse bei Verwendung der diskriminanzanalytischen Reidentifikationstechnik. Obwohl dieser Algorithmus sowohl die Möglichkeit von statistischen Doppelgängern in der Grundgesamtheit wie auch das Auftreten von Dateninkompatibilitäten berücksichtigt, steigt die Anzahl von korrekten Zuordnungen im Vergleich zu einfachen Abgleichtechniken nicht an. Ebenso wenig war es mit dieser Reidentifikationstechnik möglich, anhand der Zuordnungswahrscheinlichkeiten auch nur annähernd zwischen falschen und korrekten Zuordnungen zu trennen.

Bei der theoretischen Erörterung der verschiedenen Komponenten des Reidentifikationsrisikos hatte sich in Kapitel 4 die Schlußfolgerung ergeben, daß beim Mikrozensus und der EVS mit erhöhten Risiken gerechnet werden muß, wenn unter der Voraussetzung einer hohen Kompatiblität und eines hohen Auflösungsgrades der Überschneidungsmerkmale entweder

---

* Autoren: Uwe Blien, Heike Wirth

- bei einem Fischzugsszenario ein zugängliches Identifikationsfile existiert, oder aufgebaut werden kann, in dem die Angehörigen der Gesamtbevölkerung oder eines Teils der Bevölkerung, der durch spezifische - im Mikrodatenfile enhaltene - Merkmalsausprägungen abgrenzbar ist, vollständig oder weitgehend vollständig enthalten sind, oder

- ein Angreifer Kenntnis darüber hat, daß eine bestimmte Person im Mikrodatenfile enthalten ist (response knowledge).

Der Fall des Fischzuges wurde mit Szenario 1 überprüft, bei dem der Gelehrtenkalender eine sehr spezifische Teilpopulation weitgehend vollständig erzielt und bei dem durch regionale und sachliche Informationen ein hochauflösliches Überschneidungswissen zur Verfügung stand. Bei dieser riskanten Situation war die faktische Anonymität zweifelsfrei gegeben.

Bei der Unterstellung von Teilnahmekenntnis ergaben sich dagegen Anhaltspunkte für eine *Risikokonstellation*, bei denen unter Umständen in Einzelfällen eine erfolgreiche Reidentifikation mit vergleichsweise niedrigem Aufwand möglich erscheint. Dieser - allerdings äußerst seltene Fall, der in Angriffsszenario 4 in Abschnitt 10.2.5 diskutiert wurde, setzt das Zusammentreffen sehr spezifischer Risikofaktoren voraus und kann allgemein wie folgt charakterisiert werden:[2]

- Eine im Mikrodatenfile gesuchte Person gehört einer sehr kleinen, durch ein spezifisches Merkmal eingrenzbaren *Subpopulation* an, z.B. einer bestimmten Berufsgruppe oder Nationalität (*sachliche Tiefengliederung*);

- das Mikrodatenfile enthält *tiefgegliederte Regionalinformationen*, so daß in den jeweiligen Regionaleinheiten nur wenige Angehörige dieser spezifischen Subpopulation leben (*regionale Tiefengliederung*);

- ein Forscher, der Zugang zu den Einzelangaben des Mikrodatenfile hat, kann sich Kenntnisse über einen Angehörigen dieser spezifischen Subpopulation beschaffen und weiß, daß diese Person an der Mikrodatenerhebung, über deren Daten er verfügt, teilgenommen hat (*Teilnahmekenntnis*).

---

2 Damit es zu einem Datenangriff kommt, muß außerdem vorausgesetzt werden, daß ein Datenangreifer ein subjektives Interesse daran hat, das die denkbaren Kosten der Konsequenzen des Angriffs (Reputationsverlust, Vertragsstrafen, gesetzliche Strafen) übersteigt.

- die Merkmale der Person sind genau in der Weise im Mikrodatenfile erfaßt, wie es der Forscher vermutet (*Kompatibilität*).

Beim Zusammentreffen dieser spezifischen Bedingungen erscheint die Möglichkeit der Reidentifikation eines *Einzelfalls* ohne großen Aufwand als gegeben. Es ist wichtig darauf hinzuweisen, daß alle vier Bedingungen *gleichzeitig* erfüllt sein müssen. Bereits wenn eine der Bedingungen nicht gegeben ist, kann eine sichere Reidentifikation ohne den Aufwand unverhältnismäßig hoher Kosten nach den durchgeführten Experimenten als äußerst gering betrachtet werden. Das gleichzeitige Zusammentreffen aller Bedingungen kann bei Stichprobenerhebungen als außergewöhnlich seltenes Ereignis betrachtet werden. Dennoch sollten bei der Datenübermittlung Vorkehrungen getroffen werden, daß auch eine solche (unwahrscheinliche) Risikokonstellation ausgeschlossen ist.

Im folgenden soll daher untersucht werden, mit welchen datenorientierten Schutzmaßnahmen die oben charakterisierte Risikokonstellation entschärft und so das Risiko einer Deanonymisierung weiter verringert werden kann.

Hierfür werden in einem ersten Abschnitt zunächst allgemeine Kriterien für eine zu treffende Auswahl von Anonymisierungsmaßnahmen dargestellt. In einem weiteren Abschnitt werden ausgewählte Anonymisierungsmaßnahmen in ihrer Schutzwirkung und der damit einhergehenden Auswirkung auf das Analysepotential der Daten diskutiert. In einem dritten Abschnitt schließlich soll die von einer Substichprobenziehung und Ausprägungsvergröberung ausgehende Schutzwirkung exemplarisch, anhand des zur Verfügung stehenden Datenmaterials, überprüft werden.

11.1 Kriterien für die Auswahl von Anonymisierungsmaßnahmen

Eine Darstellung sämtlicher in der Literatur diskutierten Anonymisierungsmaßnahmen ist nicht möglich. Daher muß eine Vorauswahl getroffen werden, die sich im wesentlichen an folgenden Kriterien orientiert:

- Es müssen insbesondere einfache Abgleichtechniken durch die gewählten Maßnahmen gestört werden.

Statistisches Bundesa

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

METZLER POESCHEL