Exzerpt

Müller / Blien / Knoche / Wirth (1991): Die faktische Anonymität von Mikrodaten.

Thema: Empirische Untersuchung zur Anonymisierbarkeit von Mikrodaten unter verschiedenen Szenarios.

Dieses Exzerpt habe ich für eigene Zwecke angefertigt. Ich übernehme keine Gewähr dafür, dass es vollständig und fehlerfrei ist. Es ist nicht dazu gedacht, die Lektüre von Müller et al. (1991) zu ersetzen. Die Hervorhebungen (gelbe Markierungen) folgen subjektiven Kriterien. Kursiv gesetzte Textteile sind eigene Fragen oder Anmerkungen.

Klaudia Erhardt

Inhalt des Exzerpts

IXff Kurzfassung der Ergebnisse

XVff Ergebnisse:

- einmalige Ausprägungskombinationen sind keinesfalls schon eine hinreichende Bedingung für eine Reidentifikation. Hauptgrund: praktische Unvermeidbarkeit von Inkompatibilitäten zwischen Datenfiles.

  Abschätzung des Reidentifikationsrisikos nur durch wahrscheinlichkeitstheoretische Überlegungen führt zu erheblicher Überschätzung des realen Reidentifikationsrisikos.

- "Eine Dateninkompabilität wird um so wahrscheinlicher, je mehr Variablen benötigt werden, um eine einmalige Ausprägungskombination herzustellen. (...) Als Maßnahme zum Schutz gegen Deanonymisierung ist deshalb nicht vorrangig auf die Verhinderung einmaliger Ausprägungskombinationen abzustellen, sondern es ist eher darauf zu achten, dass keine Merkmalsausprägungen ausgewiesen werden, die so selten sind, dass durch sie allein einzelne Personen leicht identifiziert werden könnten (…)." (XVI)

- Nur diejenigen Merkmale, bei denen ein Deanonymisierungsversuch ansetzen müsste, beeinflussen das Reidentifizierungsrisiko, d.h. typische Überschneidungsmerkmale mit dem beschaffbaren Zusatzwissen. "(…) Für die Anonymisierung von Datenfiles mit Einzelangaben sind deshalb maßgeblich nur die Merkmale zu berücksichtigen, für die typischerweise Zusatzwissen vorhanden ist." (XVII)

  Oder die eine so seltene Merkmalskombination ergeben, dass durch sie einzelne Personen leicht identifiziert werden könnten.

  "Bei den vielfältigen untersuchten Datenkonstellationen und Deanonymisierungsstrategien zeigte sich, dass nur bei einer gezielten Suche oder Einzelfischzügen in besonders gelagerten, sehr seltenen Einzelfällen unter Umständen mit verhältnismäßig geringem Aufwand die Möglichkeit einer Reidentifikation besteht. Dieser Fall kann wie folgt charakterisiert werden:

  1) Eine im Mikrodatenfile gesuchte Person gehört einer sehr kleinen, durch ein spezielles Merkmal identifizierbaren Subpopulation an;

  2) Der Mikrodatenfile enthält differenzierte Regionalinformationen, so daß in den Regionaleinheiten nur wenige Personen der spezifischen Subpopulation leben;

  3) der Datenangreifer weiß, daß die gesuchte Person im Mikrodatenfile enthalten ist;

  4) die Merkmale der Person sind genau in der Weise im Mikrodatenfile erfaßt, wie es der Forscher vermutet.

  Damit es zu einem Datenangriff kommt, muss außerdem vorausgesetzt werden, daß der betreffende Forscher ein subjektives Interesse daran hat, das die denkbaren Kosten der Konsequenzen des Angriffs (…) übersteigt. Beim gleichzeitigen Zusammentreffen aller dieser speziellen Bedingungen erscheint die Möglichkeit der Reidentifikation eines Einzelfalles ohne großen Aufwand als gegeben." (XVII)

empfohlene Maßnahmen:

Die Vorschläge beziehen sich nur auf den Mikrozensus und die EVS, die Übertragbarkeit auf andere Datenbestände müsste untersucht werden (XVIII)

XX Synopse der im vorletzten Kapitel empfohlenen Maßnahmen.

Kapitel 1: Ziele u. Schwerpunkte der Untersuchung der faktischen Anonymität von Mikrodaten (1ff.)

9 "Es fehlen Kenntnisse zu den zentralen Problemen der faktischen Anonymität, vor allem ein auf realistischen Voraussetzungen folgendes Wissen über das Reidentifikationsrisiko von anonymisierten Mikrodaten, die an die Wissenschaft übermittelt werden (…)"

Kapitel 2: Stand der Forschung und gegenwärtige Praxis der Datenweitergabe (15ff)

16f. Untersuchung, welche Merkmale zu einer Identifizierung führen können, in der im Vorfeld der VZ 1987 in Hamburg durchgeführten Untersuchung "Zur Anonymität und Reidentifizierbarkeit statistischer Daten" anhand künstl. Datensatz. Folgerungen jedoch problematisch wg. Künstlichkeit des Settings

17f. Besser: AIMIPH-Projekt (Paaß/Wauschkuhn 1985)

→ es wird Zusatzwissen (Identifikationsfile) benötigt, das Überschneidungen zum Mikrodatensatz enthält, die als Schlüssel dienen können.

18 6 Szenarien im AIMIPH-Projekt untersucht, u.a. Steuerfahndung, Kripo, Journalisten, Adressverlag. Zahl der Überschneidungsmerkmale sowie ihr Diffenzierungsgrad sind bedeutsam

19 Wenn der Datenangreifer weiß, dass ein bestimmter Gesuchter an einer Erhebung teilgenommen hat, entfällt der Schutz der Stichprobeneigenschaft

21 Der Informationsgehalt der Überschneidungsmerkmale zw. Identifikationsfile und Mikrodaten ist die wesentliche Determinante für die Deanonymisierungsquote.

24 Eine Studie des US Bureau of the Census (1978) untersuchte die Praxis von Anonymisierungsmaßnahmen. Es zeigte sich, dass Anonymisierungsmaßnahmen hauptsächlich auf der Grundlage von Plausibilitätsüberlegungen u. praktischer Erfahrungen entstanden und nicht aus methodischen Modellen abgeleitet wurden.

25ff. Forschung über einzelne Anonymisierungsmaßnahmen

Kapitel 3: Methodische Grundlagen von Reidentifikationsversuchen (41ff.)

Grundprinzipien (41ff.)

42 Reidentifikation braucht Gesamtdatensatz

"Ein Reidentifikationsveruch 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"

43 Personenbezug erfordert außerdem, dass die Überschneidungsmerkmale eindeutig sind (ansonsten nur mehrdeutige Zuordnung)

44/45 Ausnahme: wenn die mehreren Personen, die in den Überschneidungsmerkmalen übereinstimmen, auch in ein oder mehreren Nutzmerkmalen identisch sind. Dann kann für eine konkrete Person nämlich ein Zusatzwissen (die für alle in Frage kommenden Personen identische Nutzmerkmale) gewonnen werden. Kommt in der Realität aber sehr selten vor.

43 Deanonymisierungsbegriff (statistical disclosure) wird teilweise auch verwendet, um Datensatzerweiterungen (Hinzuspielen weiterer Information) ohne Personenbezug zu kennzeichnen.

> "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" (44)

46 Bereits wenn entweder sowohl Mikro- wie Identifikationsfile nur eine *(jeweils eigene)* Stichprobe darstellen oder wenn die Daten nicht kompatibel sind, kann nicht mehr mit letzter Sicherheit geschlossen werden, dass die zuordenbaren Datensätze tatsächlich von der gleichen Person stammen, und nicht vielmehr statistische Doppelgänger sind.

48 Daher können in diesem Fall Zuordnungen nicht mehr mit Sicherheit, sondern nur noch mit einer gewissen Wahrscheinlichkeit vorgenommen werden.

Reidentifikationstechniken (50ff.)

Abgleich von Merkmalsausprägungen (50ff)

Kann überhaupt nur funktionieren, wenn einer der beiden Files eine Vollerhebung darstellt und die Merkmale kompatibel erhoben sind.

50/51 Der Technik wurde ein hohes Gefährdungspotential zugeschrieben und das Reidentifikationsrisiko von Mikrodaten gleich dem Anteil von unique Merkmalskombinationen angesetzt. Wurde u.a. von Brunnstein in Zus.hang mit der Volkszählung vertreten (Uniqueness-Konzept). Etwas komplexer: der Anteil von unique Merkmalskombinationen in der Grundgesamtheit wird geschätzt um darüber den Anteil von "echten" Einzelfällen in der Stichprobe zu schätzen.

Besonderheiten der einfachen Abgleichstechnik:

54/55 Die Einbeziehung weiterer Überschneidungsmerkmale erhöht nicht die Zahl möglicher Zuordnungen. Dagegen kann durch eine Verringerung der Überschneidungsmerkmale die Zahl der Zuordnungen erhöht werden.

55 Zuordnungen, die auf Basis vieler Überschneidungsmerkmale getroffen werden können, können auch auf Basis weniger Überschneidungsmerkmale noch getroffen werden, können dann aber von sehr vielen "falschen" Zuordnungen begleitet sein, von denen sie nicht unterschieden werden können. Da das Zusatzwissen des Angreifers nur einen Teil der Mikrodaten umfasst (sonst bräuchte er keine Zuordnung zu versuchen), kann er "richtige" von "falschen" Zuordnungen nicht unterscheiden.

Zuordnung durch Distanzminimierung (56ff.)

Trägt zum einen sämtliche Probleme des Abgleichs von Merkmalsausprägungen mit sich (das wären Fälle mit der Distanz 0) plus zusätzliche Unsicherheiten durch die Distanzkriterien, die angelegt werden um Identität festzustellen.

58 Kann überhaupt nur funktionieren, wenn sehr viele Überschneidungsmerkmale vorliegen und der Datenangreifer wüsste, dass eine bestimmte Person aus seinem Identifikationsfile im Mikrodatenfile vorhanden ist.

Diskriminanzanalytische Reidentifikationstechnik nach Paaß/Wauschkuhn (59ff.)

59-66 Beschreibung der Methode (Diskriminanzanalyse)

75f Bisherige Annahme: Vollerhebung. Ab dann: Einführung des Aspekts Stichprobenproblematik *(ist aber mathematisch/statistisch aufwendig nachzuvollziehen, habs daher nicht gemacht)*

79 Zusammenfassung der Technik

84f Die Diskriminanzanalytische Reidentifikationstechnik und faktische Anonymität: die Technik ist viel zu aufwendig und setzt umfassende Spezialkenntnisse sowie Spezial-Computerprogramme voraus, so dass das reale Risiko äußerst gering ist → *Frage: gilt das*

heute noch? Außerdem: mit Wahrscheinlichkeiten zu argumentieren widerspricht doch dem Personenbezug, oder?→ in Abschnitt 9.2 wird die Gefahr empirisch überprüft, mit Mikrozensusdaten und "Kürschners Deutschem Gelehrtenkalender"

Kapitel 4: Bestimmungsfaktoren des Reidentifikationsrisikos (87ff.)

87-88 Grundszenario in Anlehnung an Skinner et al. (1990) und March et al. (1991): gezielte Suche mittels einfachem Datenabgleich.
Es existiert ein Mikrodatenfile über N Fälle einer Population oder n Stichprobenfälle einer Population. Datenangreifer besitzt über 1 oder mehrere Fälle der Population Zusatzwissen über Merkmale, die im Mikrodatenfile vorkommen (= Überschneidungsmerkmale). Der Datenangreifer möchte diese Fälle im Mikrodatenfile auffinden, um sich Kenntnis über die weiteren dort enthaltenen Infos zu verschaffen. Der Datenangreifer weiß aber nicht, ob die ihn interessierenden Fälle überhaupt im Mikrodatenfile enthalten sind.
mir fällt auf: es ist hier nicht die Rede von einer Zuordnung zu konkreten Personen, sondern um Zuordnung von Fällen des Mikrodatenfiles zu Fällen des Identifikationsfiles (egal, ob Personenbezug möglich ist oder nicht).

4 Bedingungen müssen zutreffen, damit die Identifikation eines im Mikrodatenfile gesuchten Datensatzes gelingt:

A. der zu identifizierende Fall muss im Mikrodatenfile enthalten sein (Repräsentations- oder Selektivitätsproblem)

B. die Überschneidungsmerkmale müssn im Identifikationsfile und im Mikrodatenfile in identischer Weise abgebildet sein (Kompabilitätsproblem)

C. der zu identifizierende Fall muss hinsichtlich der Überschneidungsmerkmale in der Population einmalig sein (Uniqueness in der Population, nicht in der Stichprobe)

D. der Angreifer muss wissen, dass die Ausprägungskombination in der Population einmalig ist (Sicherheitsproblem).

Kapitel 4.1. Ein allgemeines Konzept für das Reidentifikationsrisiko (89ff):

Szenario der gezielten Suche (89ff)

Analyse der Wahrscheinlichkeiten für die einzelnen Bedingungen → Reidentifikationsrisiko ergibt sich aus diesen kombinierten Wahrscheinlichkeiten

90 Entscheidend für das Identifikationsrisiko ist die Einmaligkeit in der Population, nicht in der Stichprobe

> "Entscheidend für die Bestimmung des Reidentifikationsrisikos ist es deshalb zu klären, welche Anteile einmaliger Kombinationen in der Grundgesamtheit bei solchen Merkmalen und Merkmalsausprägungen bestehen, die man im Mikrodatenfile und im Identifikationsfile als vorhanden und als möglichst kompatibel abgebildet unterstellen kann" (S. 91)

Der Identifikationsfile steht für das Zusatzwissen → im Zusatzwissen und im Mikrodatenfile vorhandene Überschneidungsmerkmale!

91 In der Regel verfügt der Datenangreifer nicht über das Wissen, dass eine bestimmte Merkmalsausprägung in der Population einzigartig ist. → Eine Komponente des Reidentifikations-Risikos ist die Wahrscheinlichkeit, mit der richtig geschlossen werden kann, dass die interessierende Ausprägungskombination in der Grundgesamtheit einmalig ist. "Das Sicherheitsproblem ist gelöst, wenn eines dieser beiden Files (Mikrodaten- u. Identifikationsfile) alle Fälle der Grundgesamtheit enthält".

92/93 Das Reidentifikationsrisiko für die gezielte Suche ergibt sich aus
Der Wahrscheinlichkeit für A. mal der Wahrscheinlichkeit für B. mal der Wahrscheinlichkeit für C., gegeben B (Kompatible Abbildung in beiden Files), mal der Wahrscheinlichkeit für die Annahme der Populationsuniqueness, gegeben, dass A, B u. C erfüllt sind.

Die Buchstaben beziehen sich auf die 4 Bedingungen die auf S. 87-88 aufgelistet sind.

Fischzugsszenario (93ff)

93 Einzelfischzug: ein beliebiger Fall soll reidentifiziert werden, Massenfischzug: viele beliebige Fälle sollen reidentifiziert werden: der Angreifer geht von einzigartigen Fällen im Mikrodatenfile aus und versucht sie durch Abgleich mit dem Identifikationsfile zu identifizieren.

94 der einzige Unterschied zur gezielten Suche: An Stelle der Wahrscheinlichkeit, im Mikrodatenfile enthalten zu sein, tritt die Wahrscheinlichkeit, im Identifikationsfile enthalten zu sein. Wenn das Identifikationsfile eine (nahezu) Vollerhebung bestimmter Träger von Merkmalen ist, die im Mikrodatenfile enthalten sind (z.B. Berufe im Mikrodatenfile und ein vollständiges Ärzteverzeichnis als Id-file), dann ist die Wahrscheinlichkeit für A 1 oder nahe 1 → das Reidentifikationsrisiko ist beim Fischzugszenario höher als bei der gezielten Suche.

Response Knowledge

94-95 Wenn der Datenangreifer weiß, dass ein gesuchter Fall im Mikrodatenfile enthalten ist (response knowledge), ist die Wahrscheinlichkeit für A gleich 1. Zusätzlich kann die Wahrscheinlichkeit der Popularitätseinzigartigkeit ersetzt werden durch die Wahrscheinlichkeit einer einzigartigen Ausprägungskombination im Mikrodatenfile

95 Je kleiner die Stichprobe, desto größer das Reidentifikationsrisiko.

96 **Übersicht 4.1 stellt die Komponenten des Reidentifikationsrisikos für unterschiedliche Angriffsarten und Datengegebenheiten zusammen**

→ "Immer dann, wenn ein Angreifer bei einem der Datenfiles auf eine Totalerfassung der Population zurückgreifen kann, erhöht sich das Reidentifikationsrisiko dadurch erheblich, dass ein Angreifer durch Inspektion seiner Daten vollständige Sicherheit darüber gewinnen kann, ob in der Population eine bstimmte Ausprägungskombination einzigartig ist oder nicht. (…) Aus der Übersicht wird auch deutlich, dass Fischzugszenarien vor allem dann als mit einem höheren Risiko verbunden einzuschätzen sind, wenn der Datenangreifer auf ein Identifikationsfile zurückgreifen kann, das für ein prägendes Merkmal des gesuchten Falls (annähernd) eine Totalerhebung darstellt.

"In allen Szenarien (der Übersicht) ist die Voraussetzung der Kompatibilität der Abbildung der Fälle im Mikrodatnfile und im Identifikationsfile für die Schutzwirkung bedeutsam, auch bei Wissen um Mikrodatenfilezugehörigkeit. Außerdem muss - mit Ausnahme des Sonderfalls 'response knowledge' - immer die Bedngung der Populationseinzigartigkeit erfüllt sein. Die Möglichkeit einer Reidentifikation ist deshalb in aller Regel umso unwahrscheinlicher, je größer die Population ist, auf die bezogen ein Reidentifikationsversuch unternommen wird."

Kapitel 4.2: Einzigartigkeit als Schlüssel zur Reidentifikation (99ff.)

100 Subpopulationen (aber nur kurz am Rand erwähnt)

101 Auflösungsgrad der Merkmale: hängt mit der Zahl der Ausprägungen zusammen, und mit der Korrelation mit anderen Überschneidungsmerkmalen.

102 Datensätze mit seltenen Ausprägungskombinationen sind stärker reidentifikationsgefährdet als solche mit häufiger vorkommenden Ausprägungskombinationen.

empirische Befunde zur Wahrscheinlichkeit von Einzelfällen (102ff.)

105 Schaubild zum Zusammenhang zw. dem Anteil einzigartiger Fälle und Größe des Datenfiles (sowie Anzahl der Variablen im Datenfile)

106 Die Quote nimmt bei steigendem Populationsumfang zuerst rasch ab, dann möglicherweise asymptotisch einem Grenzwert nähernd (s. Abb. S. 105)

formale Modelle zur Schätzung der Wahrscheinlichkeit von Einzelfällen in der Population (107ff.)

damit kann das Reidentifikationsrisiko ohne Kenntnis der Grundgesamtheit abgeschätzt werden.

110 Tabelle: geschätzte Anzahl Einzelfälle (in Prozent) in der Population bei unterschiedlichen Anteilen von Einzelfällen in der Stichprobe (…)

Ergebnis:

a) wenn die Einzelfallquote in der Stichprobe max. 20% beträgt, entspricht die Einzelfallquote in der Population weitgehend der mit dem Stichprobenauswahlsatz gewichteten Quote der Stichprobe, bei größeren Einzelfallquoten in der Stichprobe nehmen die Einzelfallquoten in der Population überproportional zu. "Dennoch würden sich nach diesem Modell beachtenswerte Einzelfallquoten in der Population erst dann ergeben, wenn -- etwa beim Beispiel des Mikrozensus - in der Mikrozensus-Stichprobe Einzelfallquoten von 70 Prozent oder mehr gefunden würden. (Beispiel Mikrozensus = 1%-Stichprobe. Bei kleineren Stichproben ist die Quote deutlich geringer).

b) "dass eine bestimmte in der Stichprobe als einzigartig gefundene Ausprägungskombination auch in der Population einzigartig ist, erweist sich (…) als deutlich weniger wahrscheinlich als die Wahrscheinlichkeit von Einzelfällen in der Population insgesamt."

111 Falls diese Modell-Ergebnisse auch einer empirischen Überprüfung standhalten (…) müsste geschlossen werden, dass im Falle von Mikrozensus und EVS bei Szenarien mit gezielter Suche aus Einzigartigkeit im Mikrodatenfile praktische nie mit hinreichender Sicherheit auf Einzigartigkeit in der Population geschlossen werden kann. (…)

Nur wenn der Angreifer über ein Identifikationsfile verfügt, das die Grundgesamtheit (oder eine bestimmte Subpopulation daraus) weitgehend umfasst, könnte Wahrscheinlichkeit und Sicherheit der Populationseinzigartigkeit aus dem Identifikationsfile geschätzt werden

"Als allgemeine Schlussfolgerung ergibt sich hieraus, dass für Fälle, in denen das Mikrodatenfile eine Stichprobe mit einem Auswahlsatz von höchstens 1 Prozent ist, ein beachtenswertes Reidentifikationsrisiko allenfalls von Szenarien mit 'response knowledge' sowie von Fischzugsszenarien ausgeht, bei denen der Angreifer über ein Identifikationsfile verfügt, das annähernd die gesamte Population enthält und bei dem ein außerordentlich hoch auflösender Schlüssel von Überschneidungsmerkmalen zum Mikrodatenfile gegeben ist." (111)

•*Kapitel 4.3. Kompatibilität der Überschneidungsmerkmale (112ff.)*

112 Die Mißachtung von Dateninkompatibilität führte zu einer erheblichen Überschätzung des Reidentifikationsrisikos.

"Von Dateninkompabilität sprechen wir dann, wenn ein Fall, der gleichzeitig im Mikrodatenfile und im Identifikationsfile enthalten ist, in den entsprechenden Datensätzen dieses File in einem oder mehreren Überschneidungsmerkmalen unterschiedlich abgebildet ist." (112)

115ff **Datenfehler** können im Mikrodaten- wie im Identifikationsfile enthalten sein

121ff Mikrodaten- und Identifikationsfile können einen unterschiedlichen sachlichen und/oder zeitlichen Bezug haben: z.B. unterschiedliche Berufscodes

Der Bereich des verwendbaren Zusatzwissens verkleinert sich durch nicht zusammenfügbare Kategorien erheblich!

Die meisten Merkmale verändern sich über die Zeit → bei unterschiedlichem Datum der Information in Mikrodaten- und Identifikationsfile weichen die Überschneidungsmerkmale z.T. erheblich voneinander ab.

128 empirische Befunde zeigen, dass die Reliabilität abnimmt, je mehr Ausprägungen ein Kategorienschema hat → größere Ausdifferenzierung führt zwar zu mehr einzigartigen Merkmalskombinationen, aber gleichzeitig zu mehr Dateninkompabilität zw. Mikrodaten- und Identifikationsfile

128ff Modellrechnung zur Wahrscheinlichkeit von Dateninkompablilitäten, →

"Sie (die Modellrechnung) zeigt jedoch an, dass allein aus dem Verzeicht auf die Annahme vollständig kompatibler Daten eine erhebliche Reduzierung des Reidentifikationsrisikos resultiert." (131)

131 Folgerungen für Reidentifikationsversuche

"Die Analyse zeigt auch, dass mit der Vernachlässigung der Rolle von Dateninkompatibilitäten in der bisherigen Deanonymisierungsforschung wohl einer unrealistischen Einschätzung des Reidentifikationsrisikos Vorschub geleistet wurde."

Kapitel 5: Deanonymisierungsmotive (132ff.)

132 Der Nutzen ist als wesentlicher Teilaspekt bei der Prüfung der Unverhältnismäßigkeit des Aufwandes zu berücksichtigen.

Berufliche Motive

Aus der Handlungslogik der Sozialwissenschaft ableitbare Motive

150 Fazit: praktisch nicht vorhanden. Wenn überhaupt, dann nur als Massenfischzug (zur Ziehung einer Stichprobe). Oder zur Ergänzung eigener Surveys. In diesen Fällen müsste aber eine große Zahl erfolgreicher Deanonymisierungen stattfinden. (*Im letzten Fall müsste aber keine Deanonymisierung im eigentlichen Sinn vorgenommen werden → es geht um die Zuordnung von Datensätzen zueinander, nicht um die Zuordnung zu Personen → statistische Zwillinge?)*

→ Angriffsszenarien 1 und 3 in Kapitel 8

Außerberufliche Motive

→ Einzelfischzug bzw. gezielte Suche

"(…) so soll dennoch betont werden, daß sie [außerberufliche Motive] nicht als Hebel benutzt werden dürfen, um eine Übermittlung von Einzeldatensätzen an die Wissenschaft unter akzeptablen Bedingungen zu verhindern. Mit anderen Worten, es wäre sicherlich nicht im Sinne von § 16 Abs.6 BStatG, eine Situation zu konstruieren, bei der ein Wissenschaftler durch ein als kriminell zu bezeichnendes Verhalten (…) einen bestimmten Einzeldatensatz deanonymisiert, um dann daraus abzuleiten, daß prophylaktisch ein entsprechend hoher Anonymisierungsgrad gewährleistet sein muß, der ein derartiges Vorgehen unmöglich macht. (…) Die unterstellten Motive müssen daher einen realistischen 'Hintergrund' beibehalten und aus dem Untersuchungsgegenstand ableitbar sein." (152)

"Es kann - wie die weitere detaillierte Analyse zeigen wird - davon ausgegangen werden, daß das aus außerberuflichen Motiven resultierende Risiko einer Deanonymisierung aus Kostengründen im allgemeinen unwahrscheinlich erscheint" (154)

Weil i.d.R. andere Informationsbeschaffungswege verlässlicher und effektiver sind!

155 3 Formen werden untersucht: gezielte Suche (Beispiel: Informationen über eine bestimmte Person sollen erweitert werden), Einzelfischzug (Beispiel: ein - beliebiger - Prominenter soll identifiziert werden, um Informationen über ihn zu publizieren) und Massenfischzug

(Beispiel: in Zusammenarbeit mit einem Adresshändler sollen Informationen aus dem Mikrozensus gewonnen werden, mit denen die Adressdaten angereichert werden.)

→ Angriffsszenarien 2, 4 und 5 in Kapitel 8

Nachweis der prinzipiellen Deanonymisierbarkeit als eigenständiges wissenschaftliches Motiv (156f.)

156 Ist der Unterschied zw. absoluter und faktischer Anonymisierung. Denn hier spielt das Kostenkalkül keine Rolle. (von Lenz "Datenschutzidealist" genannt). Faktische Anonymisierung muss ein solch spezielles Deanonymisierungsrisiko in Kauf nehmen. Dieses Risiko ist demnach durch § 16 Abs. 6 BStatG abgedeckt. *Demnach auch durch § 67Abs. 8 SGB X, der Begriffsbestimmung "Anonymisierung von Sozialdaten".*

"Die Einbeziehung des Motivs des Nachweises der prinzipiellen Machbarkeit der Deanonymisierung würde die vom Gesetzgeber mit dem § 16 Abs.6 BStatG verfolgte Intention geradezu torpedieren. Mit der Regelung der faktischen Anonymität hat der Gesetzgeber ja implizit anerkannt, dass unter bestimmten Voraussetzungen Deanonymisierungen möglich sind und er bereit ist, solche Fälle als verbleibendes Risiko hinzunehmen, wenn sie nur unter Einsatz unverhältnismäßiger Mittel zustande kommen können." (157)

Kapitel 6: Zusatzwissen (158ff.)

158 Definition Zusatzwissen: die Informationen, die dem Datenangreifer über einzelne namentlich identifizierbare Personen bekannt sind und die als Überschneidungsmerkmale auch im Mikrodatenfile enthalten sind. *Namentlich identifizierbar reicht u.U. nicht aus - konkreten Personen zuordenbar. Kann auch mit der Steuernummer sein, z.B. - Namen können mehrdeutig sein: "Hans Müller" erfordert z.B. Geburtsdatum und -ort zur halbwegs sicheren Identifizierung.*

188 Fazit des Kapitels:

"Insgesamt führt die Diskussion der wichtigsten Arten personenbezogener oder personenbeziehbarer Daten, wie sie in den Sozialwissenschaften vorliegen oder zugänglich sind, zu der Einschätzung, dass damit keineswegs ein Zusatzwissen verbunden ist, von dem eine bedeutende Gefahr der Reidentifikation von Mikrodaten aus dem Mikrozensus oder der EVS ausgehen könnte. (…)

Am ehesten erwies sich noch für das Einwohnermelderegister sowie für Informationsbücher über besonders herausgehobene Personengruppen (…) daß in ihnen teilweise Informationen verfügbar sind, die für umfangreichere Deanonymisierungsversuche verwendet werden könnten. Für das Melderegister ergab sich die Vermutung, daß mit Hilfe der im Register enthaltenen Informationen zu Geschlecht, Familienstand, Geburtsdatum und Staatsangehörigkeit, Mitglieder einer bestimmten Personengruppe, nämlich ausländische Staatsangehörige, soweit eindeutig bestimmt werden könnten, daß eine Reidentifikation mit Hilfe dieser Informationen möglich erscheint. (…)" (188, 189)

191-211 Tabellen der wichtigsten Register

323 (Resumé des 6. Kapitels): Im wesentlichen können 2 Arten von Informationsquellen als Zusatzwissen für einen Deanonymisierungsversuch in Frage kommen: öffentlich zugängliche Register oder private Datenquellen einerseits, sowie sozialwissenschaftliche Datenbestände (nur Primärdaten sind personenbezogen/-beziehbar, wenn die Untersuchung selbst durchgeführt wurde).

Kapitel 7: Allgemeine Kosten-Nutzen-Überlegungen zu Deanonymisierungsversuchen (212ff.)

Subjektive Anmerkung dazu gelöscht!

Kapitel 8: Entwicklung konkreter Angriffszenarien (…) (233ff.)

235 neben den beruflichen Motiven zu Deanonymisierungsversuchen werden auch wissenschaftsfremde Deanonymisierungsversuche einbezogen, weil sie, wenn sie nur auf die Deanonymisierung einzelner Personen abzielen, vergleichsweise wenig Aufwand erfordern.

236 2 Typen von Zusatzwissen sind potentiell gefährlich:

a) bei den öffentlich oder für einen beschränkten Personenkreis zugänglichen Registern sind das berufsgruppenspezifische Adressbücher sowie das Melderegister für ausländische Staatsbürger (obwohl das Melderegister stark zugangsbeschränkt ist, so dass seine Nutzbarkeit für Identifizierungsversuche in Frage steht).

b) eigene sozialwissenschaftliche Untersuchungen von Wissenschaftlern, die gleichzeitig über die Adressdaten ihrer Probanden verfügen, weil sie die Untersuchung selbst durchführten. Ziel: Erweiterung des Merkmalspektrums

237ff Diese beiden Typen dienten für 5 hypothetische Angriffsszenarien, 2 beruflich und 3 außerberuflich motivierte. Die beiden ersten Szenarien werden empirisch, die restlichen 3 Szenarien werden argumentativ überprüft.

- Erweiterung einer eigenen Studie (über Wissenschaftler) mit Mikrozensusdaten (Szenario 1).
  Stellvertretend für die eigene Studie wird Kürschners Gelehrtenkalender benutzt.→ Identifikationsfile
  Die Motivlage ist hypothetisch, denn das Ziel würde nur ein matching erfordern, nicht eine Deanonymisierung. Abgesehen davon ist es ein realistisches und riskantes Angriffsszenario, da das Identifikationsfile nahezu als Vollerhebung vorliegt.
  Die empirische Überprüfung wird mit einer einfachen Abgleichtechnik sowie der diskriminanzanalytischen Reidentifikationstechnik vorgenommen.
- Erweiterung der Information über einen Bekannten zur Befriedigung der persönlichen Neugier (Szenario 2)
  Dieses Angriffsszenario spielt eine größere Rolle in der Diskussion. Jedoch lässt sich relativ leicht über andere Wege die Neugierde befriedigen → keine großen Antriebskräfte.
  Als Identifikationsfile dient eine sozialwissenschaftliche Erhebung mit Infos, die auch als Alltagswissen vorliegen könnten.
  Empirische Überprüfung mit einfacher Abgleichtechnik in zwei Variationen: mit und ohne response knowledge.
- Gewinnung einer Auswahlgrundlage für eine Ausländerstichprobe aus dem Mikrozensus (Szenario 3)
  Ein Wissenschaftler versucht anhand des Einwohnermelderegisters als Identifikationsfile, möglichst viele Ausländer im Mikrozensus zu identifizieren.
  Ist wissenschaftlich höchstens als Methodenvergleich zw. Mikrozensus und Studie sinnvoll.
  Da keine Aussicht besteht, ein solches Reidentifikationsexperiment auf Basis des Melderegisters *der Melderegister - Mehrzahl!* durchzuführen, wird das Szenario nur argumentativ analysiert.
  Dieses Szenario ist allerdings höchst hypothetisch: kein empirischer Forscher würde so vorgehen, um eine Ausländerstichprobe zu gewinnen, und auch nicht, um einen Methodenvergleich zum Mikrozensus vorzunehmen, alleine schon wegen der Stichprobeneigenschaft des Mikrozensus und der Ungewissheit der der eventuellen Zuordnungen.
- Gewinnung von Informationen über eine Personlichkeit des öffentlichen Lebens mit dem Ziel der anonymen Publikation (Szenario 4)
  Wird ebenfalls nur argumentativ untersucht, da eine empirische Überprüfung nicht realistisch erscheint (241)

- Gewinnung ökonomisch verwertbarer Informationen aus dem Mikrozensus mit dem Ziel des Verkaufs an einen Adressenhändler. (Szenario 5)
  Da dieses Szenarion ebenfalls eine gewisse Rolle in der Diskussion spielt, wird es einer argumentativen Analyse unterzogen
  Auch dieses Szenario stellt sich rein hypothetisch dar, denn "Insgesamt stellen Deanonymisierungen für das hier unterstellte Motiv somit schon an sich kein taugliches Mittel zur Datenbeschaffung dar" (380) *Außerdem: die Käufer von Adressen haben in der Regel flächenbezogene Aktionen vor (z.B. Direktmarketing-Aktionen) und/oder Aktionen, die sich auf eine bestimmte Gruppe beziehen, die aber möglichst weitgehend ausgeschöpft werden soll (z.B. Hocheinkommensbezieher, Rentner). Für solche Aktionen wäre die Stichprobeneigenschaft der deanonymisierten Daten ein gravierender Nachteil.*

242 Synopse der Szenarien und ihrer Randbedingungen

243f Beschreibung des Verfahrens, mit dem die Experimente unter Einhaltung des Datenschutzes durchgeführt werden können: strikte Trennung von Adressbesitzern und den Forschern, die das Experiment durchführten, sowie den Einsatz eines Datentreuhänders.

264 Erfahrungen aus der Datenaufbereitung für die Szenarien 1 und 2: Für den Gelehrtenkalender mussten die Klartext-Informationen entsprechend dem Mikrozensusvercodet werden → erhebliche Unschärfen u. z.T. Lücken. Für den bereits maschinenlesbaren Identifikationsfile für Szenario 2 (soz.wiss. Erhebung) mussten die Variablenausprägungen in zeitaufwändiger Weise in auf die Variablen des Mikrozensus gemappt werden

Kapitel 9 Empirische Überprüfung des Reidentifikationsrisikos (265ff.)

Szenario 1: Zuordnungen zw. Mikrozensus und Kürschners Gelehrtenkalender (266ff.)

Methode: einfache Zuordnung

276ff. der Erwartungswert wäre 80 korrekte Zuordnungen gewesen (Gelehrtenkalender 7.983 Fälle, Mikrozensus ist 1%-Stichprobe). Es konnten jedoch nur 14 eindeutige Zuordnungen vorgenommen werden, davon waren nur 4 korrekt.

Dagegen waren in Wirklichkeit 53 Fälle des Gelehrtenkalenders im Mikrozensus enthalten, von denen nur 4 eindeutig und 10 mehrdeutig zugeordnet werden konnten.

"Aus der Perspektive eines Angreifers erweisen sich einfache Abgleichtechniken damit als kein geeignetes Instrument für eine massenhafte Deanonymisierung von Einzeldatensätzen. (…), die unter empirischen Bedingungen allerdings immer zu erwartenden Inkompatibilitäten und statistische Doppelgänger stören diese Techniken so nachhaltig, dass auch nur annähernd sichere Zuordnungen nicht möglich sind." (282)

282 Die Wahrscheinlichkeit, dass eine eindeutige Zuordnung falsch war, lag (…) wesentlich höher als die einer korrekten Zuordnung.

Methode: diskriminanzanalytische Reidentifikationstechnik (283ff.)

300 Ergebnis nach Überprüfung durch den Treuhänder: es wurden insgesamt (in mehreren Experimenten) nur 3 verschiedene Zuordnungspaare korrekt festgestellt, die allesamt auch schon mit dem einfachen Zuordnungsverfahren gefunden worden waren. Von den Fällen, die mit mindestens 99 % Wahrscheinlichkeit von dem Verfahren einander zugeordnet worden waren, war keine einzige Zuordnung richtig.

302 Dieses Verfahren erweist sich entgegen der ursprünglichen Erwartungen nicht als leistungsfähiger als die einfache Zuordnung. Es weist zwar eine größere Anzahl von Zuordnungen aus, aber lediglich die Falschzuordnungen stiegen an.

310 Ergebnis der Tests der diskriminanzanalyitischen Reidentifikationstechnik: erstens ist das Verfahren unglaublich aufwendig *wäre das heute anders? Gibt es andere, bessere, leichter zu handhabende Verfahren?* und zweitens sind für einen Datenangreifer die wenigen korrekten Zuordnungen nicht von den viel häufigeren unkorrekten Zuordnungen zu unterscheiden.

311ff Ausblick: Auch von technischen Neuentwicklungen - wie z.B. neuronale Netze und Mustererkennung - gehen keine höheren Deanonymisierungsrisiken aus, da das Problem nicht technischer, sondern inhaltlicher Natur ist: es gibt für diese Art Anwendung prinzipiell keine Trainingsmöglichkeit (ebenso wenig wie für die Diskriminanzanalyse).

Szenario 2: Zuordnung zw. Mikrozensus und einer soz.wiss. Stichprobe mit einfacher Abgleichtechnik (323 ff.)

Angriffsszenario: gezielte Suche eines neugierigen Wissenschaftlers, der eine *bestimmte* Person aus seiner Stichprobe im Mikrozensus wiederfinden will (dieses Szenario steht gleichzeitig für den Fall des privaten Zusatzwissens (Alltagswissen) über eine Person, nur dass man diesen Fall nicht automatisch testen könnte, weil das Zusatzwissen dann uncodiert vorliegt).

Randbedingung: Keine response knowledge vorhanden (→ es ist unbekannt, ob der Gesuchte am Mikrozensus teilgenommen hat)

325 In der Literatur wird *(damals)* häufig von einem besonderen Gefährdungspotential von Haushaltskontextmerkmalen im Zusatzwissen ausgegangen.

327 Haushaltserkmale führen schnell zu einzigartigen Merkmalsausprägungen in einem Mikrodatenfile

332ff Überprüfung der Ergebnisse durch den Treuhänder: In allen 3 Versuchsanordnungen war der Anteil der korrekten Zuordnungen: 0%, der Falschzuordnungen 100%, obwohl die Einzelfälle von Versuchsphase zu Versuchsphase markant stiegen (am Ende 84% im Mikrozensus, 99% in den Studiendaten).

Es gab nur zehn Personen, die sowohl im Mikrozensus als auch in der Survey-Datei vertreten waren (N-Mikrozensus war 94.747 bzw. 53.441 in der Versuchsanordnung wo die Ein-Personen-Haushalte ausgeschlossen wurden. Im N-Survey war 2.685.)

333 Da die 10 tatsächlich in beiden Stichproben vorhandenen Fälle zu 100% nicht zugeordnet werden konnten, was ausschließlich auf Dateninkompabilität zurückzuführen ist, lautet ein wichtiges Ergebnis: "Dateninkompabilität kommt unter empirischen Bedingungen ebenso wie der Stichprobeneigenschaft von Daten eine außerordentlich wichtige Schutzfunktion bei Reidentifikationsversuchen zu."

> "Der zentrale Befund dieser Analyse ist also, daß Dateninkompatibilitäten einen impliziten Schutz vor der Möglichkeit der Reidentifikation anonymisierter Daten darstellen." (334)
>
> "Im Hinblick auf die Schutzwirkung von Inkompatibilitäten ist der folgende Sachverhalt bedeutsam: Je mehr Merkmale benötigt werden, um in einem Mikrodatenfile zu einzigartigen Ausprägungskombinationen zu gelangen, umso größer ist die Wahrscheinlichkeit, dass zumindest eines der benötigten Merkmale inkompatibel abgebildet ist. Eine korrekte Zuordnung kann jeodch nur erfolgen, wenn bei keinem einzigen Merkmal eine Inkompabilität vorliegt."

340 Die Allbus-Test-Retest-Studie von 1984, mit der untersucht wurde, wie sich die Befragten-Antworten über die Zeit ändern (3 mal die gleichen Fragebögen im Abstand von 1 Monat) kann herangezogen werden, um das Ergebnis, das auf nur 10 identischen Fällen in Mikrozensus und Survey beruht, zu verallgemeinern. → es wurde eine beträchtliche Instabilität der Antworten gefunden: nur etwa 20% der Befragten antworteten über 11 Merkmale in allen 3 Wellen gleich.

343 Üntersuchung: wenn man nur die 5 kompatibelsten Überschneidungsmerkmale nimmt, dafür aber tiefer gegliederte Regionalinformationen (hier: Regionaleinheiten von mind. 100.000 Einwohnern)

344 → Auch hier gilt: wenn ein Angreifer nicht weiß, ob sich die gesuchte Person in der Stichprobe befindet, kann eine Zuordnung nur erfolgen, wenn die zu einem Fall passenden Personen in der Grundgesamtheit ermittelt werden und anhand weiterer Informationen überprüft wird, welche Person tatsächlich im Mikrodatenfile enthalten ist.

344ff Untersuchung des Reidentifikationserfolgs bei Response Knowledge (anhand der 10 Personen, die sowohl am Survey wie am Mikrozensus teilgenommen haben). Angriffstaktik: "Hintertreppenidentifikation", d.h. sukzessive Erweiterung der einbezogenen Überschneidungsmerkmale mit daraus folgender Verkleinerung der Zahl der übereinstimmenden Datensätze in der Zieldatei.

Problem für den Datenangreifer: aufgrund von Inkompatibilitäten wird bei Einbeziehung von mehr Überschneidungsmerkmalen auch der "richtige" Fall in der Zieldatei ausgeschlossen, was der Datenangreifer so lange nicht merkt, solange es übereinstimmende Datensätze in der Zieldatei gibt. → obwohl er weiß, dass die gesuchte Person in der Zieldatei sein muss, findet er sie nicht, wenn es Inkompatibilitäten gibt.

Folgerungen für die Anonymität von Mikrodaten

"Die Zuordnungsexperimente haben gezeigt, daß die untersuchten Einelangaben faktisch anonym waren.

Der Datensatz einer beliebigen Person erwies sich als faktisch anonym auch unter der Bedingung, daß er nur insoweit anonymisiert war, als neben Name und Adresse Regionalinformationen unterhalb der Ebene der Bundesländer weggelassen wurden. (…)

Der Umfang des Zusatzwissens wirkt (…) in paradoxer Weise auf den Deanonymisierungsprozeß. Je mehr Überschneidungsmerkmale das Zusatzwissen enthält, desto höher ist der Anteil der Datensätze mit einzigartigen Ausprägungskombinationen und damit das potentielle Reidentifikationsrisiko. Mit jedem zusätzlichen Überschneidungsmerkmal erhöht sich jedoch zugleich auch die Wahrscheinlichkeit, daß ein gesuchter Fall in diesem Merkmal inkompatibel abgebildet ist, woraus Falsch- beziehungsweise Nichtzuordnungen resultieren können." (348)

"Weiterhin konnte gezeigt werden, daß Dateninkompabilitäten in einem unerwartet hohen Ausmaß auch in der bislang als höchst riskant [angesehenen] Angriffskonstellation der Teilnahmekenntnis eine immanent wichtige Schutzfunktion einnehmen. (…) Diese Ergebnisse schließen nicht aus, daß es unter der spezifischen Risikokonstellation 'Teilnahmekenntnis' mit einer gewissen Wahrscheinlichkeit möglich ist, eine gesuchte Person in einer eindeutigen Weise einem Datensatz in einer amtlichen Erhebung zuzuordnen. Solange der Angreifer seine Zuordnungen jedoch nicht verifizieren kann, und dies wäre letztendlich nur über einen Anschriftenvergleich möglich, kann er auch bei dieser speziellen Risikokonstellation nicht davon ausgehen, dass die zugeordneten Datensätze von ein- und derselben Person stammen." (349)

Argumentative Überprüfung der Szenarien

Ergeben sämtlich: Die Szenarien sind unrealistisch, weil Aufwand und Nutzen in keinem vernünftigen Verhältnis stehen oder sogar der Nutzen insgesamt fehlt → auch im Licht dieser Szenarien ist der Mikrozensus faktisch anonym.

Szenario 3: Gewinnung einer Auswahlgrundlage für eine Ausländerstichprobe (352ff.)

353 Mikrodatensatz: Mikrozensus, Identifikationsfile: Ausländer-Datensätze aus dem Melderegister

353 Überschneidungsmerkmale: Geschlecht, Geburtsjahreshälfte, Familienstand, Staatsangehörigkeit, Existenz einer Zweitwohnung, Regionalmerkmal.

Problem: schwierig, umständlich und teuer, umfangreiche Melderegister-Auszüge in zahlreichen Kommunen zu beantragen.

356ff Deanonymisierungskosten sind erheblich

362 Deanonymisierungsnutzen:

eine Stichprobe aus deanomyisierten Datensätzen des Mikrozensus als Auswahlgrundlage für eine eigene Untersuchung eignet sich wegen starker Verzerrungen nicht für wissenschaftliche Untersuchungen, da die deanonymisierbaren Daten nicht repräsentativ sind.

363 Fazit: alternative Datenbeschaffungswege (eigene Erhebung) sind für diesen Zweck geeigneter und deutlich günstiger.

Szenario 4: Gewinnung von Informationen über eine Persönlichkeit des öffentlichen Lebens mit dem Ziel der anonymen Publikation (364ff.)

Ein Mitarbeiter eines sozialwissenschaftlichen Instituts setzt sich zum Ziel, "aus dem übermittelten Mikrodatenfile der amtlicshen Statistik Datensätze über eine prominente Person herauszufischen, zu deanonymisieren und schließlich die gewonnenen Erkenntnisse anoym zu veröffentlichen".

Einzelfischzug: nicht eine bestimmte Person, sondern überhaupt eine Person soll deanonymisiert werden. Kann nur mit einer Person funktionieren, die extreme Merkmale hat.

366 Problem: ein passendes Identifikationsfile aus öffentlich zugänglichen Quellen zu finden, sowie zu erwartende Dateninkompatibilitäten. Günstiger wäre es, eine Person aus einer speziellen kleinen Personengruppe zu wählen, für die Register (Handbücher) vorliegen, etwa Abgeordnete (dann wäre Problemstellung ähnlich wie in Szenario 1)

372 Fazit: Wegen der nichtmonetären Verwertungsabsicht im Szenario dient der Vergleich der Kosten von Deanonymisierung und einer alternativen Datenbeschaffung als Maßstab.

Wenn dem Angreifer das Regionalmerkmal im Mikrozensus zur Verfügung steht, fallen u.U. nur niedrige Deanonymisierungskosten an (nicht gerechnet potentielle Sanktionen gegen den Angreifer!)

Und wo ist der Nutzen für den Angreifer im unterstellten Szenario?

Die unterstellte "anonyme Veröffentlichung" soll vermutlich die Randbedingung herstellen, dass die Sanktionen in der Kosten-Nutzen-Überlegung keine Rolle spielen. Allerdings macht diese Randbedingung das Ganze zu einem außerberuflichen Motiv, denn für einen Sozialwissenschaftler sind anonyme Veröffentlichungen aus beruflicher Sicht nutzlos.

Das Motiv kann dann nur noch sein: Jemandem aus Böswilligkeit Schaden zufügen zu wollen, falls sichergestellt ist, dass man damit nicht auffliegt.

Würde ein solches Motiv - falls es existiert - nur bei WissenschaftlerInnen vorliegen können? Nicht etwa bei MitarbeiterInnen der statistischen Ämter und Prozessdaten produzierenden Institutionen? Diese haben i.d.R. Zugang zu schwach anonymisierten, z.T. sogar zu personenbezogenen Daten!.

Szenario 5: Gewinnung ökonomisch verwertbarer Informationen zum Verkauf an einen Adresshändler (373ff.)

384 Fazit

"Wegen der wenigen Überschneidungsmerkmale sind praktisch gar keine Erfolge zu erwarten. Selbst bei ausnahmsweiser Kenntnis des Angreifers über einige, wenige Mikrozensus-Teilnahmen bliebe die Anzahl der Reidentifikationen wohl deutlich unter derjenigen, die für einen massensweisen Adressenverkauf zu Marketingzwecken notwendig wäre."

Kapitel 11 Anonymisierungsmaßnahmen (386ff.)

386ff. Theoretische Überlegungen hatten ergeben, dass beim Mikrozensus und der EVS mit erhöhten Risiken gerechnet werden muss, wenn unter Voraussetzung einer hohen Kompatibbilität und eines hohen Auflösungsgrads der Überschneidungsmerkmale entweder

- 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 enthaltene - 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).

Dieser Fall wurde mit Szenario 1 (Gelehrtenkalender) überprüft, mit dem Ergebnis, das faktische Anoynmität zweifelsfrei gegeben war.

Ergo: Alle untersuchten Szenarien ergaben die faktische Anonymität der Mikrozensusdaten bzw. der EVS.

Nur unter einer ganz speziellen Bedingungskonstellation erscheint eine Reidentifikation im Einzelfall möglich:

"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 (…) setzt das Zusammentreffen sehr spezifischer Risikofaktoren voraus und kann allgemein wie folgt charakterisiert werden:

- Die im Mikrodatenfile gesuchte Person gehört einer sehr kleinen, durch ein spezifisches Merkmal eingrenzbaren Subpopulation an (…) *(sachliche Tiefengliederung)*
- das Mikrodatenfile enthält tiefgegliederte Regionalinformationen, so daß in den jeweiligen Regionaleinheiten nur wenige Angehörige dieser spezifishen 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 (…) teilgenommen hat *(Teilnahmekenntnis)*
- Die Merkmale der Person sind genau in der Weise im Mikrodatenfile erfaßt, wie es der Forscher vermutet *(Kompatibilität)* (387)

[Außerdem weisen die Autoren in FN 2 (387) darauf hin, dass für einen Datenangriff auch noch hinzukommen muss, dass der Angreifer ein subjektives Interesse an einer Deanonymisierung hat, welches das damit verbundene Risiko (z.B. Reputationsverlust, Sanktionen) überwiegt.]

(…)

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." (387-388)

Die Wahrscheinlichkeit für die Teilnahmekenntnis in obiger Riskokonstellation verringert sich auch dadurch, dass der Forscher nicht nur Kenntnis über die Teilnahme einer beliebigen Person braucht, sondern über die Teilnahme einer Person mit extremen Merkmalen.

"Dennoch sollen bei der Datenübermittlung Vorkehrungen getroffen werden, daß auch eine solche (unwahrscheinliche) Risikokonstellation ausgeschlossen ist" (388)

→ Die weithin rezipierten und angewendeten Empfehlungen für Anonymisierungsmaßnahmen aus Müller et al. (1991) sollen also das Deanonymisierungsrisiko aus dieser, von den Autoren als äußerst unwahrscheinlich angesehenen Bedingungskonstellation auffangen (denn alle anderen untersuchten Szenarien ergaben bereits zweifelsfrei die faktische Anonymität der Mikrodaten)?.

Kriterien für die Auswahl von Maßnahmen:

- Es müssen v.a. einfache Abgleichtechniken gestört werden

  Gilt die Aussage auch heute noch?: "(…) hat gezeigt, dass bereits die Anpassung eines differenzierten statistischen Reidentifikationsalgorithmus an eine spezifische Datenbasis mit einem extrem hohen Zeit-, Kosten- und Arbeitsaufwand verbunden ist."

- Der statistische Gehalt der Daten sollte durch die gewählten Maßnahmen möglichst wenig beeinträchtigt werden. (Hinweis auf grundsätzlichen Interessen- und Zielkonflikt zw. Forschung u. Datenschutz in FN 3 (389). → Verlust an Individualinformation möglichst gering halten sowie Zusammenhänge zwischen Variablen sollten möglichst unverändert erhalten bleiben (390)

Darstellung ausgewählter Maßnahmen (391ff.)

Einbringen v. falschen Angaben (391)

Schutzwirkung: Kompabilität zu potentiellem Zusatzwissen wird reduziert

391 Einführung von Zufallsfehlern (Zufallsrauschen) durch:

a) kontinuierliche Merkmale:

- Addition von oder Multiplikation mit normalverteilten Zufallszahlen
- Zufälliges Vertauschen benachbarter Merkmalsausprägungen
- Zufälliges Runden von Merkmalsausprägungen

b) diskrete Merkmale

- Zufällige Veränderung von Merkmalsausprägungen in einem Teil der Datensätze.

Weitergehende Ansätze in Kap 2.1, Forschung: Bildung von künstlichen Datensätzen, deren statistische Eigenschaften den Originaldaten möglichst nahe kommen

Bewertung:

392: Nachhaltiger Schutz vor einfachen Abgleichtechniken, jedoch starke Verzerrungen, die die Brauchbarkeit für Analysen stark einschränken (z.T. signifikant unterschiedliche Ergebnisse in multivariaten Modellen) Gilt nur eingeschränkt für Zufallsüberlagungen von Tabellenwerten (FN 6). Außerdem für Längsschnittuntersuchungen problematisch, weil nicht unterscheidbar ist, ob Unterschiede zwischen tx und ty reale Veränderungen oder Anonymisierungsmaßnahmen sind.

393 Das Einbringen falscher Angaben ist ein viel zu schweres Geschütz, das bei der Übermittlung von faktisch anonymen Mikrodaten nicht zur Anwendung kommen sollte.

Vergröberung von Merkmalen (394)

394 Beschreibung des Risikos: gleichzeitige sachliche und regionale Tiefengliederung ermöglicht die Abgrenzung von spezifischen Einzelfällen und erleichtert die Beschaffung von Zusatzwissen.

die Risikobeschreibung stellt die fehlende Motivation unter WissenschaftlerInnen für Einzelfischzüge nicht in Rechnung!

395f **Vergröberung von Regionalmerkmalen**

"Sehr kleinräumige Regionalangaben gelten generell als anonymitätsgefährdend. Hinzu kommt, daß Regionalangaben in der Regel vergleichsweise fehlerfrei erfaßt sind. Deshalb gehört die Vergröberung des Regionalrastres international zu den Standardinstrumenten des Anonymitätsschutzes." (395)

"Wie in Kapitel 4 (…) erläutert wurde, nimmt die Zahl der Einzelfälle bei wachsender Bezugsbasis P sehr schnell ab. Ausgehend von den empirischen Ergebnissen sollte daher ähnlich wie beim Census-

Sample eine Grenze für Gebietseinheiten festgelegt werden, die im allgemeinen bei faktisch anoynmen Daten nicht unterschritten werden sollte." (396).

396 Als Kompromiss zw. Analysepotential und Anonymisierung: Entweder regionale Tiefengliederung oder sachliche Tiefengliederung, jedoch nicht beides zusammen.

396f **Vergröberung von schwach besetzten oder extremen Merkmalswerten**

"Ausgehend von einem festzulegenden Besetzungsminimum können alle Merkmalsausprägungen, die unterhalb dieser Schwelle liegen, nach inhaltlichen Kriterien mit anderen Ausprägungen zusammengefaßt werden. Auf diese Weise kann ausgeschlossen werden, daß eine durch ein Merkmal eindeutig charakterisierte Person in dem Datenmaterial enthalten ist. (..)."

??? vorher wurde doch ausführlich dargestellt, dass das Deanonymisierungsrisiko nicht mit Uniqueness im Mikrodatenfile, sondern mit Uniqueness in der Grundgesamtheit einhergeht, und außerdem mit dem Wissen des Datenangreifers, dass in der Grundgesamtheit Uniqueness vorliegt ??

Antwort daruf: damit wird das Risiko eines Einzelangriffs bei response knowledge abgefangen.

Maßnahmen ohne direkte Datenmodifikation (398)

Die wichtigste und wirkungsvollste, dabei die wissenschaftliche Verwendung am wenigsten störende Maßnahme ist die Ziehung einer Substichprobe (→ 400ff.)

Kapitel 14: Empfehlungen für die Weitergabe von Einzelangaben aus dem Mikrozensus und der EVS an die Wissenschaft (440ff.)

→ s. dazu: Unterlage zum Workshop: "Synopse der Anonymisierungsmaßnahmen"

Begründung der Empfehlungen (445ff.)

hier nur selektiv dargestellt.

446ff Begründung in Bezug auf eine Reidentifikation durch Massenfischzug

447 Staatsangehörigkeit wird als "ein in der Regel leicht und kompatibel mit dem Mikrodatenfile erfahrbares Merkmal" angesehen, → Weitergabe nur stark vergröbert

Wie wäre denn die Staatsangehörigkeit für einen Massenfischzug leicht zu erfahren? Über die Einwohnermeldeämter? Über das Ausländer-Zentralregister?

449 ff Begründung in Bezug auf Reidentifikation durch gezielte Suche:

451-452 Es wird erneut mit Nachdruck betont, dass die Konstellation, unter der es durch gezielte Suche zu Deanonymisierungen kommen kann, äußerst selten ist

Insgesamt fällt auf, dass die Autoren widersprüchlich argumentieren: einerseits hat die Untesuchung ergeben, dass Mikrozensus und EVS faktisch anonym sind, mit Ausnahme einer speziellen und äußerst selten vorkommenden Bedingungskonstellation, die auf die Wissenschaft nicht stärker zutrifft als auf die Mitarbeiter der Datenproduzenten.

Andererseits werden Empfehlungen für Maßnahmen gegeben, mit denen der bereits faktisch anonyme Datensatz noch weiter gegen Deanonymisierungsversuche geschützt werden soll, im Grunde unter

Berufung auf wissenschaftsfremde Motive und Konstellationen, und unter der Annahme von response knowledge

Die Anonymisierungsmaßnahmen sind jedenfalls mit einem dicken "Sicherheitspuffer" berechnet.