Dr. Rainer Lenz, Dr. Daniel Vorgrimler

Geheimhaltungsmethoden auf dem Prüfstand – eine Analyse anhand der Umsatzsteuerstatistik

In der wissenschaftlichen Diskussion zur Geheimhaltungsproblematik von Einzeldaten ist bisher relativ wenig über die Möglichkeit bekannt, unter realitätsnahen Bedingungen anonymisierten Merkmalsträgern mit Hilfe von so genanntem Zusatzwissen direkte Identifikatoren korrekt zuzuordnen. Ebenso ist unzureichend geklärt, in welchem Umfang Geheimhaltungsmethoden unter realen Bedingungen in der Lage sind, richtige Zuordnungen zu verhindern. Realitätsnah bedeutet dabei, Einzeldaten einer externen Datei als Zusatzwissen zu verwenden und dieses mit einem verfügbaren Zuordnungsalgorithmus einer anonymisierten Zieldatei zuzuordnen.

Der vorliegende Beitrag nimmt sich dieser beiden Fragestellungen an und will mit seinen Resultaten die Diskussion zu den Fragen weiter voranbringen. Hierzu wird im Rahmen von Simulationsexperimenten eine externe Datei als Zusatzwissen einem Teilbereich der Umsatzsteuerstatistik 2000 zugeordnet. Die verschiedenen Ergebnisse zeigen, welche Chancen bestehen, Merkmalsträger zu „reidentifizieren" bzw. ob die Daten durch einen „natürlichen Schutz" bereits so sicher sind, dass die statistische Geheimhaltung bereits durch „formale" Anonymisierung gewahrt bleibt. Verwendet wird dazu ein Zuordnungsalgorithmus, der im Rahmen von CASC – einem europäischen Forschungsprojekt zur Geheimhaltung – entwickelt wurde. Dieser soll noch in diesem Jahr allgemein zugänglich gemacht werden. Die wiederholte Durchführung der Experimente mit anonymisierten Dateien zeigt anschließend die Effektivität verschiedener Geheimhaltungsmethoden bei der Verhinderung von richtigen Zuordnungen auf. Neben traditionellen Anonymisierungsmaßnahmen steht dabei die Anonymisierung durch Mikroaggregation im Mittelpunkt des Interesses.

Vorbemerkung

Empirische Wirtschaftsforscher klagen zunehmend über einen mangelnden Zugang zu wirtschaftsstatistischen Einzeldaten der amtlichen Statistik. Durch diesen sei das Potenzial an möglichen Analysen, die mit amtlichen Erhebungen denkbar sind, nur ungenügend ausgeschöpft. Dabei wurde bereits 1987 im Bundesstatistikgesetz[1]) mit dem § 16 Abs. 6 der Wissenschaft ein privilegierter Zugang zu Mikrodaten der amtlichen Statistik eingeräumt. Dieser Paragraph erlaubt die Übermittlung von Einzeldaten an die Wissenschaft, sofern diese nur mit unverhältnismäßig hohem Aufwand reidentifiziert werden können. „Unverhältnismäßig" bedeutet, dass die Kosten einer Reidentifikation deren Nutzen übersteigen (faktische Anonymität).

Während die Bereitstellung von faktisch anonymisierten Personendaten bereits seit längerem bewährte Praxis ist, ergeben sich für wirtschaftsstatistische Einzeldaten besondere Probleme, die bisher eine faktische Anonymisierung unmöglich erscheinen ließen. Verglichen mit Personenerhebungen liegen bei Unternehmenserhebungen wesentlich kleinere Grundgesamtheiten zugrunde, sodass die Besetzungszahlen einzelner Gruppen häufig kleiner sind. Die Verteilungen der quantitativen Merkmale sind wesentlich heterogener und es treten mit größerer Wahrscheinlichkeit

1) Gesetz über die Statistik für Bundeszwecke (Bundesstatistikgesetz – BStatG) vom 22. Januar 1987 (BGBl. I S. 462, 565), zuletzt geändert durch Artikel 16 des Gesetzes vom 21. August 2002 (BGBl. I S. 3322).

dominierende Fälle auf. Des Weiteren sind die Stichprobenauswahlsätze bei Unternehmenserhebungen wesentlich höher als bei Personenerhebungen. Publizitätspflichten der Unternehmen einerseits und kommerzielle Datenbanken andererseits bewirken, dass einem Externen, der Einzeldaten ihrem Träger zuordnen will, bei Wirtschaftsdaten besseres Zusatzwissen zur Verfügung steht als bei Personendaten. Schließlich wird der Nutzen aus der Kenntnis von Unternehmens- und Betriebsdaten wesentlich höher eingestuft als von Daten aus Personen- und Haushaltserhebungen. Ein rationaler Datenangreifer wird deshalb auch höhere Kosten für eine Deanonymisierung akzeptieren.

Diese Besonderheiten wirtschaftsstatistischer Einzeldaten stellen besondere Ansprüche an die verwendeten Anonymisierungsmethoden. Um diesen Ansprüchen gerecht zu werden, entwickelte die Anonymisierungsforschung neue Verfahren, welche die Vertraulichkeit der Daten sichern sollen. Im folgenden Beitrag wird mit der Mikroaggregation ein Verfahren auf den Prüfstand gestellt, welches in der internationalen Literatur größere Bedeutung erlangt hat.[2]) Verglichen wird dabei die Wirkung des Verfahrens sowohl in Kombination mit verschiedenen traditionellen Anonymisierungsmaßnahmen als auch mit einer rein traditionellen Anonymisierung. Darüber hinaus wird untersucht, ob eine formale Anonymisierung (d. h. Löschung der direkten Identifikatoren) bereits genügend Schutz bietet.

Diese Fragen werden mit Hilfe von Simulationen analysiert. Es wird einerseits versucht, über so genannte Massenfischzüge[3]) möglichst viele Merkmalsträger (in diesem Falle Unternehmen) und andererseits über Einzelangriffe bestimmte Merkmalsträger zu reidentifizieren. Das grundsätzliche Vorgehen und die Beschreibung des hierzu verwendeten Algorithmus ist Thema der folgenden Kapitel 1 und 2. Die Ergebnisse der eigentlichen Simulationen stehen in Kapitel 3 im Mittelpunkt. Hier werden zunächst die Zieldaten und das so genannte Zusatzwissen beschrieben sowie die Höhe des „natürlichen Schutzes“ der Daten gegen Reidentifikationsversuche bestimmt. Anschließend werden in Abschnitt 3.2 die Ergebnisse der Simulationen vorgestellt. Ein Fazit in Kapitel 4 schließt den Beitrag ab.

1 Die Grundvoraussetzungen eines Datenangriffes

Um Merkmalsträger einer vertraulichen Datei erfolgreich reidentifizieren zu können, sind folgende Grundannahmen für einen Datenangreifer nötig[4]):

- Zusatzwissen über die gesuchten Merkmalsträger (etwa in Form einer externen Unternehmensdatenbank)
- Kenntnis über die Teilnahme des gesuchten Merkmalsträgers an der Erhebung (Zieldaten)
- Merkmale, welche sowohl in externen als auch in Zieldaten enthalten sind (Überschneidungsmerkmale).

Darüber hinaus muss der Datenangreifer persönlich von der Richtigkeit der Zuordnung überzeugt sein, was bei einem Massenfischzug nahezu unmöglich erscheint.

Im Folgenden seien $A=\{a_1,\ldots,a_m\}$ und $B=\{b_1,\ldots,b_n\}$ Mengen von Merkmalsträgern der externen Daten und der Zieldaten, welche eine nichtleere Menge von Überschneidungsmerkmalen teilen. Die Menge der Überschneidungsmerkmale werde mit $\{v_1,\ldots,v_k\}$ bezeichnet. Wir zerlegen die Überschneidungsmerkmale in zwei Klassen, *kategoriale* und *metrische* Merkmale. Bei kategorialen Merkmalen wird zwischen *nominalen* Merkmalen (es gibt keine lineare Ordnung zwischen den Kategorien) und *ordinalen* Merkmalen (die Kategorien besitzen eine lineare Ordnung, wobei Differenzen zwischen Kategorien keinen Sinn ergeben) unterschieden. Unter metrischen Merkmalen versteht man Merkmale, bei denen eine Differenzbildung von Ausprägungen Bedeutung hat, wie zum Beispiel „Größe“ und „Gewicht“ einer Person oder in unserem Falle „Umsatz“ und „Anzahl der Beschäftigten“ eines Unternehmens. Offenbar kann jedes metrische Merkmal über Klassenbildung des Wertebereiches in ein kategoriales Merkmal umgewandelt werden (man betrachte etwa Umsatz- oder Beschäftigtengrößenklassen).

Bei der Simulation eines Datenangriffes sei nun $(a,b)\in A\times B$ ein Kandidatenpaar für eine mögliche Zuordnung. Wir fordern, dass beide Merkmalsträger a und b in zuvor festgelegten Überschneidungsmerkmalen übereinstimmen. Solche Merkmale werden *Blockmerkmale* genannt, da sie den gesamten Datenbestand in überschneidungsfreie (disjunkte) Blöcke unterteilen. Es erscheint vernünftig, solche Merkmale als Blockmerkmale zu bestimmen, welche einerseits die Merkmalsträger gut charakterisieren und andererseits wenig fehlerbehaftet sind. Letzteres gilt bei unseren Betrachtungen insbesondere für einige kategoriale Merkmale, die nicht durch das auf die Zieldaten angewendete Anonymisierungsverfahren verändert wurden. In der Praxis können allerdings auch in solchen Merkmalen signifikante Abweichungen zwischen verschiedenen Erhebungen auftreten (siehe Abschnitt 3.1.1).

2 Der Zuordnungsalgorithmus

Ein potenzieller Datenangreifer steht vor dem Entscheidungsproblem, ob ein Paar $(a,b)\in A\times B$ von Merkmalsträgern zu demselben zugrunde liegenden Individuum (oder Unternehmen) gehört. Hierzu ist ein vernünftiger Ähnlichkeitsbegriff notwendig. Grob gesprochen tritt die größtmögliche für den Datenangreifer messbare Ähnlichkeit zwischen zwei Merkmalsträgern dann auf, wenn sie in sämtlichen

2) Stellvertretend siehe Domingo-Ferrer, J./Mateo-Sanz, J. M.: “Practical data-oriented microaggregation for statistical disclosure control”, IEEE Transactions on Knowledge and Data Engineering, Vol. 14(1), 2002, S. 189 ff.

3) Bei einem Massenfischzug versucht ein Datenangreifer mit Hilfe einer externen Datenbank als Zusatzwissen soviel Merkmalsträger der Zieldaten wie möglich zu identifizieren. Im Gegensatz zum Einzelangriff ist er dabei nicht an einem speziellen Merkmalsträger interessiert.

4) Siehe Brand, R./Bender, S./Kohaut, S.: “Possibilities for the creation of a scientific-use-file for the IAB-establishment-panel”, Proceedings of the Joint Eurostat/UN-ECE Work Session on Statistical Data Protection, Thessaloniki, 1999, S. 57 ff.

Überschneidungsmerkmalen übereinstimmen. Da im Falle geringer Abweichungen in den Überschneidungsmerkmalen die beiden betrachteten Merkmalsträger als stark verwandt empfunden werden, hängt die Qualität der Zuordnung offenbar im Wesentlichen von der Wahl des Ähnlichkeitsmaßes ab. Wir führen hierzu folgende Distanztypen ein:

Sei k die Anzahl der Überschneidungsmerkmale. Für das metrische Merkmal v_i und das Paar (a, b) von Merkmalsträgern wird die i-te Komponentendistanzfunktion definiert durch $d_i(a,b) = (a^{(i)} - b^{(i)})^2$, wobei $a = (a^{(1)}, \ldots, a^{(k)})$ und $b = (b^{(1)}, \ldots, b^{(k)})$ die auf die k Überschneidungsmerkmale reduzierten Merkmalsträger sind. Weiterhin definieren wir mit

$$d_i(a,b) := \begin{cases} 0, \text{wenn } a^{(i)} = b^{(i)} \\ 1 \text{ andernfalls} \end{cases}$$

eine Komponentendistanzfunktion für nominale Merkmale v_i und mit

$$d_i(a,b) := \frac{|\{c_j \mid \min(a^{(i)}, b^{(i)}) \leq_i c_j <_i \max(a^{(i)}, b^{(i)})\}|}{r}$$

eine Komponentendistanzfunktion für ordinale Merkmale v_i, ausgenommen Blockmerkmale, wobei $c_1 <_i c_2 <_i \ldots <_i c_r$ der geordnete Wertebereich des Merkmals v_i ist. Um Skalierungsprobleme zu vermeiden, werden die berechneten Komponentendistanzen normiert mittels der bewährten max-min-Standardisierung

$$\tilde{d}_r(a,b) := \frac{d_r(a,b) - \min_{(\alpha,\beta)\in A\times B} d_r(\alpha,\beta)}{\max_{(\alpha,\beta)\in A\times B} d_r(\alpha,\beta) - \min_{(\alpha,\beta)\in A\times B} d_r(\alpha,\beta)}.$$

Es sei im Weiteren NV die Indexmenge der metrischen und CV die Indexmenge der kategorialen Merkmale. Die Gesamtdistanz ergibt sich als gewichtete Summe über alle Komponentendistanzen:

$$\begin{aligned} d(a,b) &:= \sum_{i\in NV} \lambda_i \tilde{d}_i(a,b) + \sum_{i\in CV} \lambda_i \tilde{d}_i(a,b) \\ &= \sum_{i\in NV} \lambda_i \tilde{d}_i(a,b) + \sum_{i\in C_{ord}} \lambda_i \tilde{d}_i(a,b) + \sum_{i\in C_{nom}} \lambda_i \tilde{d}_i(a,b), \end{aligned}$$

wobei der Distanzbeitrag der kategorialen Merkmale gemäß der vorherigen Überlegungen in einen ordinalen (C_{ord}) und einen nominalen (C_{nom}) Anteil zerfällt. Es ist sinnvoll, die Blockung der Daten ebenfalls in die Distanzberechnung einzubeziehen[5]), da das in Komplexitätsuntersuchungen gewöhnlich nicht berücksichtigte blockweise Einlesen der Daten sehr zeitaufwändig sein kann.

Wir sind nun in der Lage, die einzelnen Schritte des Zuordnungsalgorithmus zu skizzieren:

1. Input: $T = (\tau_1, \ldots, \tau_k)$ Vektor der Merkmalstypen,

   $\Lambda = (\lambda_1, \ldots, \lambda_k)$ Vektor der Merkmalsgewichte,

   $A = \{a_1, \ldots, a_m\}$ und $B = \{b_1, \ldots, b_n\}$ wie in Kapitel 1 beschrieben.

   Dabei ist T ein k-Tupel ganzer Zahlen mit den folgenden Eigenschaften:

   $\tau_i = 0$ genau dann, wenn v_i das Identifikatormerkmal (z. B. Unternehmensregisternummer),

   $\tau_i = 1$ genau dann, wenn v_i ein Blockmerkmal,

   $\tau_i = 2$ genau dann, wenn v_i ein ordinales Merkmal und

   $\tau_i = 3$ genau dann, wenn v_i ein nominales Merkmal ist.

   Andernfalls wird v_i als metrisches Merkmal angenommen.

   Mit Λ wird ein mit T verträgliches k-Tupel reeller Zahlen bezeichnet. Das heißt, $\tau_i = 0$ impliziert $\lambda_i = 0$ und $\tau_i = 1$ impliziert $\lambda_i = 1$. Die verbleibenden Gewichte werden geeignet standardisiert, sodass die Gleichung $\Sigma_{i=1}^{k} \lambda_i = 1$ erfüllt ist.

2. Berechnung der Gesamtdistanzen $d_{ij} := d(a_i, b_j)$ für $i = 1, \ldots, m$ und $j = 1, \ldots, n$.

3. Aufsteigende Sortierung der Gesamtdistanzen in eine Liste L.

4. Solange L nicht leer ist:

   Betrachte das erste Element d_{ij} in L und ordne (a_i, b_j) zu.

   Entferne alle Elemente d_{rs}, für welche $r = i$ oder $s = j$ gilt.

5. Output: Relative Häufigkeit der korrekten Zuordnungen.

3 Anwendung auf reale Daten

Dieses Kapitel enthält die Ergebnisse der Anwendung des oben beschriebenen Algorithmus auf reale Daten. Die Massenfischzüge werden sowohl mit formal anonymisierten Daten (Weglassen direkter Identifikatoren wie Name und Adresse) als auch mit Daten, die über unterschiedliche Methoden probeanonymisiert wurden, durchgeführt.

3.1 Die Zieldaten und das Zusatzwissen

Als Zieldaten für den Massenfischzug wurden die Daten der Umsatzsteuerstatistik 2000 verwendet. Die Umsatzsteuerstatistik dient der Beurteilung der Struktur und Wirkungsweise der Umsatzsteuer und ihrer wirtschaftlichen Bedeutung. Aus der Beobachtung der Umsätze ergeben sich wertvolle Informationen für die Haushaltsplanungen und Steuerschätzungen des Bundes, der Länder und der Gemeinden. Die Umsatzsteuerstatistik ist nicht allein ein Instrument der Fiskal- und Steuerpolitik; sie dient darüber hinaus auch der allgemeinen Wirtschaftsbeobachtung. Mit ihren Angaben über die Entwicklung der Umsätze in allen Bereichen der Volkswirtschaft liefert sie Informationen, die in dieser Vollständigkeit in keiner anderen Bundesstatistik enthalten sind. Die Ergebnisse der Umsatzsteuerstatistik sind eine wichtige Datenbasis für die Erstellung der Volkswirtschaftlichen Gesamtrechnungen. Aufgrund ihrer tiefen

5) Siehe Lenz, R.: "Disclosure of confidential information by means of multi objective optimisation", Comparative analysis of enterprise (micro) data conference, London 2003.

wirtschaftssystematischen Gliederung lassen sich mit Hilfe der Umsatzsteuerstatistik auch branchenspezifische Analysen sowie Konzentrationsuntersuchungen durchführen.

Die Umsatzsteuerstatistik ist eine Sekundärstatistik, die auf die Daten zurückgreift, die bei der Finanzverwaltung anfallen. Erfasst werden alle Unternehmen, die Umsatzsteuer-Voranmeldungen abgeben, mit einem Jahresumsatz (ohne Umsatzsteuer) von über 32 500 DM (16 617 Euro). In der Umsatzsteuerstatistik für das Jahr 2000 sind rund 2,9 Mill. Unternehmen erfasst.

Für die Durchführung der Reidentifikationsexperimente wurden die Merkmale Wirtschaftszweignummer, Umsatz, Rechtsform und Regionalkennung[6]) als Schlüssel- bzw. Überschneidungsmerkmale verwendet.

Das verwendete Zusatzwissen enthielt knapp 9 300 Unternehmen mit 20 oder mehr Beschäftigten aus den Abteilungen 10 bis 37 (Abschnitt C „Bergbau und Gewinnung von Steinen und Erden“ sowie D „Verarbeitendes Gewerbe“) der Klassifikation der Wirtschaftszweige, Ausgabe 1993 (WZ 93). Rund 37 000 Unternehmen wiesen seitens der Umsatzsteuerstatistik diese Charakteristika auf, sodass diese prinzipiell als gesuchte Unternehmen in Frage kamen. Daraus ergab sich das Ziel, die 9 300 Unternehmen, für die Zusatzwissen vorlag, innerhalb der 37 000 Unternehmen der Zieldaten zu reidentifizieren (siehe die Übersicht „Prinzip des Massenfischzuges“).

Prinzip des Massenfischzuges

<table>
<tr><td colspan="2">Zusatzwissen (9 300 Unternehmen)</td><td></td></tr>
<tr><td>Identifikatoren (Name, Anschrift)</td><td>Umsatz, Rechtsform, Regionalkennung, Wirtschaftszweignummer</td><td></td></tr>
<tr><td></td><td>Umsatz, Rechtsform, Regionalkennung, Wirtschaftszweignummer</td><td>Zielmerkmale (z.B. Umsatzwachstum 1999 bis 2000)</td></tr>
<tr><td></td><td colspan="2">Zieldaten (37 000 Unternehmen)</td></tr>
</table>

In Anlehnung an Höhne, J./Sturm, R./Vorgrimler, D.: „Konzept zur Beurteilung der Schutzwirkung von faktischer Anonymisierung“ in WiSta 4/2003, S. 288.

Für die Massenfischzüge wurden die kategorialen Merkmale unterschiedlich anonymisiert. So wurden Massenfischzüge auf der Ebene des WZ-Vier- bis Einstellers[7]) durchgeführt und es wurden Massenfischzüge simuliert, bei denen ganz auf eine Untergliederung der Abschnitte C und D verzichtet wurde. Die Rechtsform wurde einmal mit acht Ausprägungen und einmal mit vier Ausprägungen verwendet. Der Regionalschlüssel wies neun, sieben oder drei Ausprägungen auf.

Die verwendeten Umsatzsteuerstatistikdaten wurden darüber hinaus nach vier unterschiedlichen Verfahren anonymisiert:

– Das erste Verfahren war die formale Anonymisierung als schwächste Anonymisierungsform. Dabei werden lediglich die direkten Identifikatoren (Name, Anschrift) gelöscht.

– Die zweite Anonymisierung beschränkte sich auf traditionelle Anonymisierungsmaßnahmen wie zum Beispiel Vergröberungen von Merkmalen (siehe Abschnitt 3.1.2).

– Das dritte Verfahren war die schwächste Form der Anonymisierung durch Mikroaggregation. Dabei wird jedes stetige Merkmal separat mikroaggregiert (siehe Abschnitt 3.1.3).

– Die vierte Art beinhaltete dagegen die stärkste Form der Mikroaggregation, bei der sämtliche stetigen Merkmale gemeinsam behandelt werden (siehe Abschnitt 3.1.3).

3.1.1 Natürlicher Schutz der Zieldaten

In diesem Beitrag werden zwei Arten des Schutzes analysiert: zum einen der durch die Anonymisierungsmaßnahme bewirkte Schutz und zum anderen der Schutz, der dadurch entsteht, dass die Merkmalsausprägungen bereits zwischen formal anonymisierten Zieldaten und Zusatzwissen für einen Merkmalsträger erheblich voneinander abweichen können. Dies kann man als natürlichen Schutz der Daten gegen Reidentifikation bezeichnen. Tabelle 1 zeigt den Grad der Übereinstimmung zwischen beiden Erhebungen für die Merkmale Umsatz und Wirtschaftsklassifikation. Es wird beispielsweise ersichtlich, dass bei fast 30% (20%) aller

Tabelle 1: Übereinstimmungen in den Merkmalsausprägungen zwischen Zieldaten und Zusatzwissen

Gegenstand der Nachweisung	Unternehmen	
	Anzahl	%
mit Abweichungen in der Ausprägung des Merkmals Umsatz		
Abweichung geringer als ...		
1%	3 546	*38,2*
5%	6 541	*70,5*
10%	7 539	*81,2*
25%	8 395	*90,4*
50%	8 706	*93,8*
Insgesamt ...	9 283	*100*
mit identischer WZ 93[1])-Klassifizierung auf der Ebene der		
4-Steller	5 206	*56,1*
3-Steller	5 917	*63,7*
2-Steller	7 007	*74,5*
1-Steller	7 823	*84,3*
Insgesamt ...	9 283	*100*

1) Klassifikation der Wirtschaftszweige, Ausgabe 1993. Als 1-Steller wird hier die Zehnerstelle des Codes für die Abteilungen (Zweisteller) bezeichnet.

6) Als Regionalkennung wurden die siedlungsstrukturellen Kreistypen verwendet. Diese nichtadministrativen Schlüssel dienen dem intraregionalen Vergleich. Es wird nach „Kernstädten“ und sonstigen Kreisen bzw. Kreisregionen unterschieden. Als Kernstädte werden kreisfreie Städte mit mehr als 100 000 Einwohnern ausgewiesen. Kreisfreie Städte unterhalb dieser Größe werden mit ihrem Umland zu Kreisregionen zusammengefasst. Die Typisierung der Kreise und Kreisregionen erfolgt außerhalb der Kernstädte nach der Bevölkerungsdichte. Um den großräumigen Kontext zu berücksichtigen, wird dann weiter nach der Lage im siedlungsstrukturellen Regionstyp differenziert. Mit dieser Einordnung wird der Überlegung Rechnung getragen, dass die Lebensbedingungen in den Kreisen sowie ihre Entwicklung wesentlich auch von der Entwicklung und der Struktur der jeweiligen Region bzw. des Regionstyps abhängig sind. Insgesamt ergeben sich neun Kreistypen, die im Folgenden unter dem Merkmal BBR9 angegeben werden. Nach einer Zusammenfassung auf regionale Grundtypen ergeben sich drei Ausprägungen, die im Folgenden unter dem Merkmal BBR3 ausgewiesen werden (siehe Bundesamt für Bauordnung und Raumwesen, *www.bbr.bund.de/raumordnung/raumbeobachtung/gebietstypen2.htm*).

7) Als „Einsteller“ wird hier die Zehnerstelle des (zweistelligen) Codes für die Abteilungen in der WZ 93 bezeichnet (mit den Ausprägungen 1, 2 und 3).

Unternehmen der Umsatzwert im Zusatzwissen um mindestens 5% (10%) vom Umsatzwert in der Umsatzsteuerstatistik abweicht. Des Weiteren zeigt die Tabelle, dass nur 56% aller Unternehmen die gleiche vierstellige Wirtschaftszweignummer in Zusatzwissen und Zieldaten aufweisen.

Darüber hinaus ist noch von Bedeutung, dass etwa 2% der Unternehmen Unterschiede in der Regionalkennung aufweisen. Aufgrund dieser zum Teil mangelnden Übereinstimmung ist es von vornherein unmöglich, alle im Zusatzwissen vorhandenen Unternehmen richtig zuzuordnen bzw. zu reidentifizieren. Die aus diesem Grund nicht auffindbaren Unternehmen sind durch den natürlichen Schutz bereits ausreichend geschützt. Tabelle 2 enthält die Anzahl der geschützten und ungeschützten Unternehmen in Abhängigkeit von der verwendeten Gliederungstiefe der Wirtschaftszweigklassifikation. Es zeigt sich, dass beispielsweise bei einem Reidentifikationsversuch, der auf der Ebene der Viersteller durchgeführt wird, rund 45% aller Unternehmen von vornherein geschützt sind (37% bei Verwendung der Dreisteller der Wirtschaftszweigklassifikation usw.).

Tabelle 2: Durch Abweichungen geschützte und ungeschützte Unternehmen

Auf Ebene der WZ 93[1])	Ungeschützte Unternehmen		Geschützte Unternehmen	
	Anzahl	%	Anzahl	%
4-Steller	5 120	*55,2*	4 163	*44,8*
3-Steller	5 812	*62,6*	3 471	*37,4*
2-Steller	6 878	*74,1*	2 405	*25,9*
1-Steller	7 673	*82,7*	1 610	*17,3*
0-Steller	9 097	*98,0*	186	*2,0*
Insgesamt[2]) ...	9 283	*100*	0	*0,0*

1) Klassifikation der Wirtschaftszweige, Ausgabe 1993. Als 1-Steller wird hier die Zehnerstelle des (zweistelligen) Codes für die Abteilungen bezeichnet; der 0-Steller steht für die einbezogenen Abschnitte C „Bergbau und Gewinnung von Steinen und Erden“ und D „Verarbeitendes Gewerbe“ ohne weitere Untergliederung. – 2) Ohne Abweichungen im Regionalschlüssel.

3.1.2 Traditionelle Anonymisierung

Bei einer der betrachteten Probeanonymisierungen der Zieldaten wurden die Merkmalsträger durch traditionelle Methoden anonymisiert. Dabei wurden folgende Maßnahmen durchgeführt:

- Kürzung der Wirtschaftszweignummer auf zwei Stellen (Abteilungen),
- jede Abteilung musste mit mindestens 3 500 Unternehmen besetzt sein, weshalb kleinere Abteilungen zusammengefasst wurden,
- Vergröberung der Rechtsform auf vier Ausprägungen (anstelle von acht),
- Topcoding des Umsatzes unter Berücksichtigung zweier Abschneidegrenzen (500 Mill. Euro und 1 Mrd. Euro Umsatz). Die Umsätze der Unternehmen oberhalb der Abschneidegrenze wurden durch den Durchschnitt der Umsätze aller Unternehmen oberhalb der Abschneidegrenze ersetzt (Replacement),
- Runden des Umsatzes bei Unternehmen mit bis zu 500 Mill. Euro Umsatz.

3.1.3 Anonymisierung durch Mikroaggregation

Erfahrungen früherer Untersuchungen anhand der Kostenstrukturerhebung zeigen, dass es mit Hilfe der (mehrdimensionalen) Mikroaggregation möglich erscheint, anonymisierte Daten zu erstellen, die genügend Analysepotenzial beinhalten[8]). Die Mikroaggregation unterteilt zunächst die metrischen Merkmale in Gruppen, wobei anzuraten ist, hoch korrelierte Merkmale zusammen zu gruppieren. Innerhalb der Gruppen werden die Merkmale standardisiert und für jeden Merkmalsträger aufsummiert, sodass die Merkmalsträger nach diesen so genannten Z-scores geordnet werden können. Im nächsten Schritt werden für eine vorgegebene ganze Zahl k (hier k = 3) die Merkmalsträger mit dem größten bzw. kleinsten Z-score mit ihren k – 1 nächsten Nachbarn (bei mehrdimensionaler Mikroaggregation bezüglich der euklidischen Norm) zusammengelegt und bei den so entstandenen k-Tupeln (hier: Tripeln) merkmalsweise die Werte gemittelt.

Um eine Ober- und eine Untergrenze für das mit der Mikroaggregation verbundene Reidentifikationsrisiko bestimmen zu können, werden im nächsten Abschnitt die schwächste und die stärkste Variante der Mikroaggregation auf die Zieldaten angewendet. Während bei der schwächsten Variante (MA21G) jedes metrische Merkmal seine eigene Gruppe definiert, werden bei der stärksten Variante (MA1G) alle metrischen Merkmale zusammen gruppiert, sodass Tripel von Merkmalsträgern erzeugt werden, welche in allen metrischen Merkmalen übereinstimmen und sich lediglich in den kategorialen Merkmalen (wie z. B. Wirtschaftszweigzuordnung oder Regionalkennung) unterscheiden können. In der Tat kann man bei der Variante MA1G große Abweichungen der anonymisierten von den Originaldaten beobachten. Mehr als 60% der veränderten Werte weichen um mehr als 10% von ihrem Originalwert ab. Bei der Variante MA21G hingegen werden die Originaldaten nur sehr geringfügig modifiziert. Hier weichen mehr als 99,9% der veränderten Werte um weniger als 5% von ihren Originalwerten ab. Dies gilt sogar für beinahe 90% der Unternehmen mit mehr als 500 Beschäftigten, welche bekanntermaßen besonders reidentifikationsgefährdet sind.

3.2 Der Massenfischzug

Bei den nun folgenden Ergebnissen der Massenfischzüge werden zwei Arten von Trefferquoten unterschieden: Bei der ersten Art werden die Treffer in Beziehung zu allen gesuchten Unternehmen gesetzt. Bei der zweiten Art dagegen werden die Treffer lediglich im Verhältnis zu den Unternehmen betrachtet, die nicht aufgrund des natürlichen Schutzes per se sicher sind (siehe Tabelle 2). Diese Trefferquote wird im Folgenden als „korrigierte Trefferquote“ bezeichnet. Als Blockmerkmale wurden die kategorialen Merkmale „Rechtsform“, „Wirtschaftszweigzuordnung“ und „Regionalken-

8) Siehe Rosemann, M./Vorgrimler, D./Lenz, R.: „Erste Ergebnisse faktischer Anonymisierung wirtschaftsstatistischer Einzeldaten“ in Allgemeines Statistisches Archiv, Heft 1, 2004, S. 73 ff.

nung“ verwendet. Als metrisches Überschneidungsmerkmal stand das Merkmal „Umsatz“ zur Verfügung.

3.2.1 Ergebnisse in Abhängigkeit von der Tiefe der wirtschaftlichen Gliederung

Tabelle 3 zeigt die Trefferquoten in Abhängigkeit von der Tiefe der wirtschaftlichen Gliederung bis hin zum Verzicht auf eine Untergliederung der einbezogenen Abschnitte der WZ 93 (hier als „0-Steller“ bezeichnet). Letzteres kommt der Annahme gleich, dass ein Datenangreifer keinerlei Wissen über die ökonomischen Aktivitäten des gesuchten Unternehmens hat, außer dass es sich um ein Unternehmen des Bergbaus und der Gewinnung von Steinen und Erden oder des Verarbeitenden Gewerbes handelt.

Es wird offensichtlich, dass die schwächste Form der Mikroaggregation (MA21G) die Trefferquoten nicht reduzieren kann.[9]) Die Unterschiede zwischen den Erhebungen erschweren einen Datenangriff bereits so stark, dass eine zusätzliche minimale Veränderung der Werte durch MA21G bei einem Massenfischzug nicht ins Gewicht fällt. Betrachtet man die Ergebnisse des Matches auf Basis der 0-Steller, so erkennt man sogar einen enthüllenden Effekt der Mikroaggregation. Es wurden mehr Unternehmen nach der Anonymisierung durch MA21G gefunden als bei lediglich formaler Anonymisierung (1 270 gegenüber 1 259). MA1G dagegen generiert nahezu sichere Daten. Allerdings sei an dieser Stelle darauf hingewiesen, dass diese Form der Mikroaggregation das Analysepotenzial stark einschränkt.[10]) Die Trefferquoten bei Verwendung der vierstelligen und der dreistelligen Wirtschaftszweignummer ähneln sich bei den formal und den durch MA21G anonymisierten Daten sehr. Beim Übergang von vier auf drei Stellen werden zwar die Blöcke, innerhalb derer versucht wird die Unternehmen zu reidentifizieren, größer (und damit auch das Risiko der Falschzuordnung, siehe Tabelle 5), andererseits nimmt die Anzahl derjenigen Unternehmen ab, die durch den natürlichen Schutz geschützt sind. Diese beiden gegenläufigen Effekte heben sich in diesem Fallbeispiel auf. Wird die Wirtschaftszweignummer aber weiter gekürzt, so kann ein Anstieg der Schutzwirkung beobachtet werden. Bei der traditionellen Anonymisierung zeigt sich, dass die großen Unternehmen hierbei besser geschützt werden als bei den anderen Verfahren. Dies liegt darin begründet, dass die Anonymisierungsmethoden in diesem Fall besonders auf die großen Unternehmen ausgerichtet waren (z. B. Topcodingverfahren beim Umsatz).

Tabelle 4 zeigt die korrigierten Trefferquoten. Diese Raten zeigen die reine Schutzwirkung aufgrund der getroffenen Anonymisierungsmaßnahmen für die metrischen Merkmale. Es zeigt sich einerseits wiederum, dass durch die Variante MA21G die Trefferquote nicht wesentlich reduziert wird. Andererseits wird in der Tabelle die Schutzwirkung einer Vergröberung der Gliederung nach Wirtschaftszweigen deutlich, da dieser Effekt nicht mehr durch einen Verlust an natürlichem Schutz konterkariert wird. Daher ist erkennbar, dass die Trefferquoten beim Übergang vom Vierauf den Dreisteller der WZ 93 zurückgehen. Die erste Zeile der Tabelle zeigt die Effektivität des verwendeten Matchingverfahrens. Ohne Inkompatibilitäten bei den Blockmerkmalen und ohne zusätzliche Anonymisierung kann der Algorithmus drei von vier Unternehmen richtig zuordnen. Bei den größten Unternehmen wurden sogar sämtliche Unternehmen richtig reidentifiziert.

Tabelle 3: Korrekt zugeordnete Unternehmen

Zieldaten	WZ 93[1])	Insgesamt		Beschäftigtengrößenklasse[2])											
				1		2		3		4		5		6	
		Anzahl	%	Anzahl	%	Anzahl	%	Anzahl	%	Anzahl	%	Anzahl	%	Anzahl	%
Formal anonymisiert	4-Steller	3 726	40,1	188	35,3	1 764	35,7	1 583	45,7	169	54,9	15	57,7	7	70
	3-Steller	3 720	40,1	189	35,5	1 781	36,1	1 565	45,1	162	52,6	17	65,4	6	60
	2-Steller	3 287	35,4	168	31,6	1 557	31,5	1 371	39,5	167	54,2	16	61,5	8	80
	1-Steller	1 951	21,0	95	17,9	912	18,5	800	23,1	131	42,5	9	34,6	4	40
	0-Steller	1 259	13,6	61	11,5	586	11,9	508	14,7	89	28,9	11	42,3	4	40
MA21G[3])	4-Steller	3 723	40,1	188	35,3	1 763	35,7	1 580	45,6	170	55,2	15	57,7	7	70
	3-Steller	3 709	39,9	192	36,1	1 785	36,1	1 545	44,6	164	53,3	15	57,7	8	80
	2-Steller	3 282	35,4	168	31,6	1 558	31,5	1 362	39,3	170	55,2	16	61,5	8	80
	1-Steller	1 934	20,8	94	17,7	906	18,3	790	22,8	130	42,2	9	34,6	5	50
	0-Steller	1 270	13,7	60	11,3	593	12,0	505	14,5	100	32,5	8	30,8	4	40
MA1G[4])	4-Steller	2 593	27,9	114	21,4	1 076	21,8	1 214	35,0	166	53,9	17	65,4	6	60
	3-Steller	2 169	23,4	84	15,8	853	17,3	1 043	30,1	162	52,6	19	73,1	8	80
	2-Steller	1 332	14,4	37	6,9	471	9,5	656	18,9	144	46,8	18	69,2	6	60
	1-Steller	497	5,4	11	2,1	165	3,3	244	7,0	65	21,2	9	34,6	3	30
	0-Steller	241	2,6	5	0,9	73	1,5	117	3,4	36	11,7	7	26,9	3	30
Traditionell anonymisiert		2 792	30,0	142	26,7	1 335	27,0	1 181	34,0	127	41,2	5	19,2	2	20

1) Klassifikation der Wirtschaftszweige, Ausgabe 1993. Als 1-Steller wird hier die Zehnerstelle des (zweistelligen) Codes für die Abteilungen bezeichnet; der 0-Steller steht für die einbezogenen Abschnitte C „Bergbau und Gewinnung von Steinen und Erden“ und D „Verarbeitendes Gewerbe“ ohne weitere Untergliederung. – 2) 1 = weniger als 25 Beschäftigte, 2 = 25 bis unter 100 Beschäftigte, 3 = 100 bis unter 1 000 Beschäftigte, 4 = 1 000 bis unter 5 000 Beschäftigte, 5 = 5 000 bis unter 15 000 Beschäftigte, 6 = 15 000 und mehr Beschäftigte. – 3) Schwächste Variante der Mikroaggregation. – 4) Stärkste Variante der Mikroaggregation.

9) Die Reduzierung der Trefferquoten ist aber nur einer der beiden in Höhne u. a. beschriebenen Aspekte der Schutzwirkung faktischer Anonymisierung. Eine Analyse, die beide Aspekte der Schutzwirkung berücksichtigt, beinhaltet der Beitrag von Lenz, R./Sturm, R./Vorgrimler, D.: „Maße für die faktische Anonymität von Mikrodaten“ auf S. 621 ff. in diesem Heft.

10) Zu den Auswirkungen der Anonymisierung auf das Analysepotenzial siehe Rosemann, M.: „Erste Ergebnisse von vergleichenden Untersuchungen mit anonymisierten und nicht anonymisierten Einzeldaten am Beispiel der Kostenstrukturerhebung und der Umsatzsteuerstatistik“ in Gnoss, R./Ronning, G. (Hrsg.): „Anonymisierung wirtschaftsstatistischer Einzeldaten“, Band 42 der Schriftenreihe „Forum der Bundesstatistik“, Wiesbaden 2003, S. 154 ff.

Tabelle 4: Korrigierte Trefferquoten
Prozent

Zieldaten	WZ 93[1]	Insgesamt	Beschäftigtengrößenklasse[2]					
			1	2	3	4	5	6
Formal anonymisiert	4-Steller	72,8	70,7	70,0	75,1	84,9	83,3	100
	3-Steller	64,0	62,0	61,0	66,7	76,4	81,0	75,0
	2-Steller	47,8	46,3	44,4	50,2	68,4	72,7	100
	1-Steller	25,4	22,8	23,0	26,7	49,1	39,1	50,0
	0-Steller	13,8	11,7	12,1	15,0	29,7	42,3	44,4
MA21G[3]	4-Steller	72,7	70,7	69,9	75,0	85,4	83,3	100
	3-Steller	63,8	62,9	61,2	65,8	77,4	71,4	100
	2-Steller	47,7	46,3	44,4	49,9	69,7	72,7	100
	1-Steller	25,2	22,5	22,9	26,4	48,7	36,1	62,5
	0-Steller	14,0	11,5	12,2	14,9	33,3	30,8	44,4
MA1G[4]	4-Steller	50,6	42,9	42,7	57,6	83,4	94,4	85,7
	3-Steller	37,3	27,5	29,2	44,4	76,4	90,5	100
	2-Steller	19,4	10,2	13,4	24,0	59,0	81,8	75,0
	1-Steller	6,5	2,6	4,2	8,2	24,4	39,1	37,5
	0-Steller	2,7	1,0	1,5	3,5	12,0	26,9	33,3
Traditionell anonymisiert		40,6	39,1	38,0	43,2	52,1	22,7	25,0

1) Klassifikation der Wirtschaftszweige, Ausgabe 1993. Als 1-Steller wird hier die Zehnerstelle des (zweistelligen) Codes für die Abteilungen bezeichnet; der 0-Steller steht für die einbezogenen Abschnitte C „Bergbau und Gewinnung von Steinen und Erden“ und D „Verarbeitendes Gewerbe“ ohne weitere Untergliederung. – 2) 1 = weniger als 25 Beschäftigte, 2 = 25 bis unter 100 Beschäftigte, 3 = 100 bis unter 1 000 Beschäftigte, 4 = 1 000 bis unter 5 000 Beschäftigte, 5 = 5 000 bis unter 15 000 Beschäftigte, 6 = 15 000 und mehr Beschäftigte. – 3) Schwächste Variante der Mikroaggregation. – 4) Stärkste Variante der Mikroaggregation.

Tabelle 5 zeigt, dass je größer einzelne Datenblöcke sind (d. h. je mehr Unternehmen eine bestimmte Kombination aus Rechtsform, Regionalkennung und Wirtschaftszweigzuordnung erfüllen), desto geringer ist die Wahrscheinlichkeit, ein Unternehmen richtig zuordnen zu können. Dies spricht dafür, eine Mindestanzahl von Unternehmen je Gliederungsebene der Wirtschaftszweigklassifikation nicht zu unterschreiten.

Tabelle 5: Korrelation zwischen Trefferquoten und Besetzungszahlen der Wirtschaftszweige

Zieldaten	WZ 93[1]		
	4-Steller	3-Steller	2-Steller
Formal anonymisiert	–0,504	–0,683	–0,777
MA21G[2]	–0,499	–0,665	–0,779
MA1G[3]	–0,474	–0,644	–0,661
Traditionell anonymisiert			–0,413

1) Klassifikation der Wirtschaftszweige, Ausgabe 1993. – 2) Schwächste Variante der Mikroaggregation. – 3) Stärkste Variante der Mikroaggregation.

Tabelle 6 enthält ausgewählte beschreibende Statistiken der richtig zugeordneten Unternehmen. Zum Vergleich zeigt die letzte Zeile die Statistik aller Unternehmen des Zusatzwissens. Man erkennt, dass mit steigendem Anonymisierungsgrad die durchschnittliche Zahl der Beschäftigten der jeweils richtig zugeordneten Unternehmen steigt. Eine Ausnahme stellt die traditionelle Anonymisierung dar. Es zeigt sich daher auch bei dieser Betrachtung, dass mit Ausnahme der traditionellen Anonymisierung die Anonymisierungsmaßnahmen stärker bei den kleinen Unternehmen als bei den großen wirken. Betrachtet man allerdings die Größe des jeweils kleinsten gefundenen Unternehmens, dann erkennt man, dass auch die kleinsten Unternehmen richtig zugeordnet werden können.

Tabelle 6: Beschreibende Statistik der reidentifizierten Unternehmen (Zahl der Beschäftigten)

Zieldaten	WZ 93[1]	Identifikation des größten Unternehmens	Durchschnittliche Größe	Kleinstes Unternehmen	Standardabweichung	Unternehmen
			Beschäftigte			Anzahl
Formal anonymisiert	4-Steller	ja	431,3	20	4 326,3	3 726
	3-Steller	ja	421,8	20	4 279,4	3 720
	2-Steller	ja	488,6	20	4 673,4	3 287
	1-Steller	nein	445,5	20	5 469,0	1 951
	0-Steller	nein	626,0	20	6 897,7	1 259
MA21G[2]	4-Steller	ja	433,2	20	4 328,9	3 723
	3-Steller	ja	443,3	20	4 396,3	3 709
	2-Steller	ja	491,5	20	4 677,9	3 282
	1-Steller	ja	532,4	20	8 924,5	1 934
	0-Steller	ja	644,5	20	5 883,1	1 270
MA1G[3]	4-Steller	ja	561,3	20	5 117,5	2 593
	3-Steller	ja	688,4	20	5 743,1	2 169
	2-Steller	ja	936,6	20	7 126,3	1 332
	1-Steller	ja	1 200,6	21	8 924,5	497
	0-Steller	ja	1 946,6	20	12 750,1	241
Traditionell anonymisiert		nein	305,6	20	2 279,2	2 792
Alle Unternehmen des Zusatzwissens		–	295,8	20	2 888,7	9 283

1) Klassifikation der Wirtschaftszweige, Ausgabe 1993. Als 1-Steller wird hier die Zehnerstelle des (zweistelligen) Codes für die Abteilungen bezeichnet; der 0-Steller steht für die einbezogenen Abschnitte C „Bergbau und Gewinnung von Steinen und Erden“ und D „Verarbeitendes Gewerbe“ ohne weitere Untergliederung. – 2) Schwächste Variante der Mikroaggregation. – 3) Stärkste Variante der Mikroaggregation.

3.2.2 Effekte durch die Vergröberung der Rechtsform

Im nächsten Schritt wurde die Rechtsform wie bei der traditionellen Anonymisierung auf vier Kategorien vergröbert und die Massenfischzüge bei den ansonsten formal anonymisierten und bei den mit der Mikroaggregation anonymisierten Daten wiederholt.[11] Im Folgenden wird die Schutzwirkung dieser Vergröberung beschrieben.

11) Eine Wiederholung der Massenfischzüge bei den mit traditionellen Maßnahmen anonymisierten Merkmalsträgern erübrigt sich, da dort diese Maßnahme von vornherein angewandt wurde. Daher enthält die Tabelle 7 auch keine Ergebnisse für die traditionelle Anonymisierung.

Tabelle 7: Korrekt zugeordnete Unternehmen absolut nach Vergröberung der Rechtsform auf vier Kategorien

Zieldaten	WZ 93[1]	Insgesamt		Beschäftigtengrößenklasse[2]											
				1		2		3		4		5		6	
		Anzahl	%	Anzahl	%	Anzahl	%	Anzahl	%	Anzahl	%	Anzahl	%	Anzahl	%
Formal anonymisiert	4-Steller	3 527	*37,9*	175	*32,9*	1 644	*33,3*	1 526	*44,0*	163	*52,9*	13	*50,0*	6	*60*
	3-Steller	3 501	*37,7*	175	*32,9*	1 656	*33,5*	1 492	*43,0*	156	*50,7*	15	*70,0*	7	*70*
	2-Steller	2 945	*31,7*	146	*27,4*	1 389	*28,1*	1 244	*35,9*	151	*49,0*	12	*46,1*	3	*30*
	1-Steller	1 587	*17,1*	75	*14,4*	718	*14,5*	669	*19,3*	115	*37,3*	8	*30,8*	2	*20*
	0-Steller	1 081	*11,4*	51	*9,6*	488	*9,9*	441	*12,7*	92	*12,7*	7	*26,9*	2	*20*
MA21G[3]	4-Steller	3 527	*37,9*	175	*32,9*	1 645	*33,3*	1 525	*44,0*	163	*52,9*	13	*50,0*	6	*60*
	3-Steller	3 499	*37,7*	175	*32,9*	1 656	*33,5*	1 491	*43,0*	115	*50,3*	15	*57,7*	7	*70*
	2-Steller	2 941	*31,7*	146	*27,4*	1 393	*28,2*	1 236	*35,6*	151	*49,0*	11	*42,3*	4	*40*
	1-Steller	1 569	*16,9*	76	*14,3*	712	*14,4*	660	*19,0*	108	*35,1*	8	*30,8*	5	*50*
	0-Steller	1 055	*11,4*	52	*9,8*	484	*9,8*	426	*12,3*	81	*26,3*	7	*26,9*	5	*50*
MA1G[4]	4-Steller	2 340	*25,2*	100	*18,8*	919	*18,6*	1 143	*32,9*	158	*51,3*	15	*57,7*	5	*50*
	3-Steller	1 881	*20,3*	69	*13,0*	690	*14,0*	957	*27,6*	141	*45,8*	17	*65,4*	7	*70*
	2-Steller	1 037	*11,2*	23	*4,3*	355	*15,1*	525	*15,1*	116	*37,7*	11	*42,3*	7	*70*
	1-Steller	317	*3,4*	7	*1,3*	89	*1,8*	160	*4,6*	49	*15,9*	6	*23,1*	6	*60*
	0-Steller	199	*2,1*	7	*1,3*	72	*1,5*	85	*2,5*	28	*9,1*	3	*11,5*	4	*40*

1) Klassifikation der Wirtschaftszweige, Ausgabe 1993. Als 1-Steller wird hier die Zehnerstelle des (zweistelligen) Codes für die Abteilungen bezeichnet; der 0-Steller steht für die einbezogenen Abschnitte C „Bergbau und Gewinnung von Steinen und Erden“ und D „Verarbeitendes Gewerbe“ ohne weitere Untergliederung. – 2) 1 = weniger als 25 Beschäftigte, 2 = 25 bis unter 100 Beschäftigte, 3 = 100 bis unter 1 000 Beschäftigte, 4 = 1 000 bis unter 5 000 Beschäftigte, 5 = 5 000 bis unter 15 000 Beschäftigte, 6 = 15 000 und mehr Beschäftigte. – 3) Schwächste Variante der Mikroaggregation. – 4) Stärkste Variante der Mikroaggregation.

Der Schutzeffekt, der mit einer Vergröberung der Rechtsform einhergeht, ist deutlich schwächer ausgefallen als erwartet. Das kommt unerwartet, da die Rechtsform in beiden Erhebungen für die gemeinsamen Merkmalsträger identisch ausgewiesen war. Die Vergröberung konnte daher ihre volle Schutzwirkung entfalten und wurde nicht – wie bei der Vergröberung der wirtschaftlichen Gliederung – durch einen geringeren natürlichen Schutz konterkariert. Der Vergleich der Tabellen 3 und 7 verdeutlicht darüber hinaus die Beobachtung, dass eine Verringerung der Tiefe der wirtschaftlichen Gliederung eine höhere Schutzwirkung als die vorgeschlagene Vergröberung der Rechtsform bewirkt.

Die durch die Vergröberung der Rechtsform einhergehende leichte Verstärkung der Anonymisierung hat wiederum zur Folge, dass die durchschnittliche Größe der richtig zugeordneten Unternehmen ansteigt. Auch hier werden demnach besonders die kleineren Unternehmen geschützt.

3.2.3 Effekte durch die Vergröberung der Regionalkennung

Neben der Vergröberung der Rechtsform stellt die Vergröberung der Regionalkennung eine weitere Anonymisierungsmaßnahme dar. Als Regionalkennung wurde wie erwähnt der siedlungsstrukturelle Kreistyp verwendet. Dieser besteht aus einer Hauptstufe mit drei Ausprägungen und einer Unterstufe, durch die der BBR9[12]) seine insgesamt neun Ausprägungen erhält. Durch den Verzicht auf die Unterstufe reduziert sich die Anzahl der Ausprägungen auf drei. Im Folgenden wird daher vom BBR3 gesprochen, wenn auf die Unterstufe verzichtet wird. Alternativ könnte auch der regionsstrukturelle Kreistyp verwendet werden. Dieser besitzt sieben Ausprägungen, also nur zwei weniger als der BBR9. Die Schutzwirkung, die durch eine solche „Vergröberung“ entsteht, ist allerdings vernachlässigbar, sodass auf eine weitere Diskussion auf Basis des BBR7 verzichtet werden kann.

Im Gegensatz zur Vergröberung der Rechtsform wird die Vergröberung auf den BBR3 auch auf die traditionelle Anonymisierung angewendet. Tabelle 8 enthält die Ergebnisse der Massenfischzüge.

Mit der Vergröberung der Regionalkennung wird eine größere Anonymisierungswirkung erzielt als mit der vorhergehenden Vergröberung der Rechtsform. Dies verdeutlicht die Bedeutung von Regionalkennungen bei einer versuchten Deanonymisierung. Ein Vergleich von Tabelle 3 und Tabelle 8 zeigt, dass eine Anonymisierung durch Reduzierung der Tiefe der wirtschaftlichen Gliederung von Vierstellern auf Zweisteller der Klassifikation der Wirtschaftszweige sowie durch Vergröberung der Rechtsform und der Regionalkennung zu einem Rückgang der Trefferquoten von teilweise mehr als 40% führt, was zeigt, dass traditionelle Maßnahmen bei der Anonymisierung wirtschaftsstatistischer Einzeldaten eine wichtige Rolle spielen können. Dies deuteten bereits die Ergebnisse an, die bei den Massenfischzügen mit den traditionell anonymisierten Daten erzielt wurden. Allerdings sei darauf hingewiesen, dass die Vergröberung der Regionalkennung bei der traditionellen Anonymisierung bei den großen Unternehmen sogar eine enthüllende Wirkung hatte, das heißt es konnten nach der Vergröberung mehr Unternehmen reidentifiziert werden. Demnach kann bei solchen Simulationen auch der Zufall eine Rolle spielen.

3.3 Der Einzelangriff

Das Ziel eines Einzelangriffs ist die Gewinnung von Informationen über einen spezifischen Merkmalsträger. Dabei sammelt ein Datenangreifer aus verschiedenen externen

12) Zur Definition siehe Fußnote 6.

Tabelle 8: Korrekt zugeordnete Unternehmen nach Vergröberung der Rechtsform und der Regionalkennung

Zieldaten	WZ 93[1]	Insgesamt		Beschäftigtengrößenklasse[2]											
				1		2		3		4		5		6	
		Anzahl	%	Anzahl	%	Anzahl	%	Anzahl	%	Anzahl	%	Anzahl	%	Anzahl	%
Formal anonymisiert	4-Steller	3 114	33,6	150	28,2	1 425	28,9	1 365	39,4	153	49,7	14	53,9	7	70,0
	3-Steller	2 900	31,2	145	27,3	1 334	27,0	1 253	36,1	150	48,7	12	46,2	6	60,0
	2-Steller	2 100	22,6	95	17,9	980	19,8	896	25,8	117	37,9	10	38,5	2	20,0
	1-Steller	889	9,6	44	8,3	386	7,8	376	10,8	74	24,0	7	26,9	2	20,0
	0-Steller	534	5,8	30	5,6	239	4,8	214	6,2	45	14,6	5	19,2	1	10,0
MA21G[3]	4-Steller	3 112	33,5	150	28,2	1 425	28,9	1 361	39,2	155	50,3	14	53,9	7	70,0
	3-Steller	2 904	31,3	146	27,4	1 339	27,1	1 251	36,1	150	48,7	12	46,2	6	60,0
	2-Steller	2 105	22,7	96	18,1	991	20,1	890	25,7	113	36,7	10	38,5	5	50,0
	1-Steller	897	9,7	42	7,9	396	8,0	374	10,8	70	22,7	9	34,6	6	60,0
	0-Steller	511	5,5	30	5,6	232	4,7	199	5,7	41	13,3	6	23,1	3	30,0
MA1G[4]	4-Steller	1 706	18,4	63	11,8	599	12,1	882	25,4	142	46,1	14	53,9	6	60,0
	3-Steller	1 250	13,5	43	8,1	400	8,1	655	18,9	129	41,9	15	57,7	8	80,0
	2-Steller	557	6,0	13	2,4	171	3,5	282	8,1	73	23,7	11	42,3	7	70,0
	1-Steller	139	1,5	2	0,4	40	0,8	63	1,8	24	7,8	4	15,4	6	60,0
	0-Steller	69	0,7	3	0,6	19	0,4	28	0,8	11	3,6	5	19,2	3	30,0
Traditionell anonymisiert		2 081	22,2	94	16,7	982	19,9	883	25,5	111	36,0	6	23,1	5	50,0

1) Klassifikation der Wirtschaftszweige, Ausgabe 1993. Als 1-Steller wird hier die Zehnerstelle des (zweistelligen) Codes für die Abteilungen bezeichnet; der 0-Steller steht für die einbezogenen Abschnitte C „Bergbau und Gewinnung von Steinen und Erden“ und D „Verarbeitendes Gewerbe“ ohne weitere Untergliederung. – 2) 1 = weniger als 25 Beschäftigte, 2 = 25 bis unter 100 Beschäftigte, 3 = 100 bis unter 1 000 Beschäftigte, 4 = 1 000 bis unter 5 000 Beschäftigte, 5 = 5 000 bis unter 15 000 Beschäftigte, 6 = 15 000 und mehr Beschäftigte. – 3) Schwächste Variante der Mikroaggregation. – 4) Stärkste Variante der Mikroaggregation.

Quellen Informationen über das gesuchte Individuum bzw. Unternehmen und versucht anschließend, mit diesem Wissen den gesuchten Merkmalsträger in den Zieldaten zu reidentifizieren. Ein solcher „Angriff“ wurde anhand der Umsatzsteuerstatistik simuliert. Dabei wurde versucht 15 Unternehmen zu reidentifizieren, die in der Erhebung lediglich formal anonymisiert enthalten sind. Die Überschneidungsmerkmale waren die Regionalkennung, die Wirtschaftszweigzuordnung, die Rechtsform und die Umsätze der Jahre 2000 und 1999 (die Überschneidungsmerkmale standen aber nicht in jedem Fall zur Verfügung). Mit Hilfe dieser Überschneidungsmerkmale gelang es 6 der 15 Unternehmen eindeutig und richtig zu identifizieren.

Wie beim Massenfischzug besteht die Hauptfehlerquelle darin, dass die Ausprägungen des Zusatzwissens sehr deutlich von den Ausprägungen der Umsatzsteuerstatistik abweichen können, die Merkmalsträger daher bereits natürlich geschützt sind. Darüber hinaus kann man sogar feststellen, dass im Gegensatz zu anderen Statistiken[13]) die Umsatzsteuerstatistik dem Datenangreifer im Rahmen eines Einzelangriffs keine zusätzlichen Überschneidungsmerkmale gegenüber einem Massenfischzug bietet. Aus diesem Grund ist das Risiko der Reidentifikation für ein spezifisches Unternehmen der Umsatzsteuerstatistik in einem Einzelangriffsszenario nicht höher zu bewerten als im Szenario des Massenfischzugs.

4 Fazit und Ausblick

Im vorliegenden Beitrag wurden verschiedene Methoden der Anonymisierung wirtschaftsstatistischer Einzeldaten dahingehend getestet, inwieweit sie geeignet erscheinen, Reidentifikationen zu verhindern. Hierzu wurde ein Algorithmus gewählt, der im Rahmen des EU-Projektes CASC (Computational Aspects of Statistical Confidentiality) entwickelt wurde und in näherer Zukunft allgemein zugänglich gemacht werden soll.[14]) Die Leistungsfähigkeit des gewählten Algorithmus zeigte sich bei Zuordnungsversuchen, bei denen auf Anonymisierungen verzichtet wurde und bei denen diejenigen Unternehmen nicht betrachtet wurden, die aufgrund von Dateninkompatibilitäten bereits einen natürlichen Schutz genießen. Dass dieser natürliche Schutz bereits einen wesentlichen Beitrag zur Erreichung faktischer Anonymität darstellt, konnte ebenfalls gezeigt werden.

Eine Anonymisierung der Merkmalsträger durch Vergröberung der diskreten Merkmale wird teilweise durch die Aufhebung der Abweichungen zwischen Zusatzwissen und Zieldaten – wodurch der natürliche Schutz der Daten sinkt – konterkariert. So erweist sich eine Verringerung der Gliederungstiefe von vier auf drei Stellen der Klassifikation der Wirtschaftszweige als nicht sonderlich hilfreich. Erst die Reduzierung auf zwei Stellen führt zu einem signifikanten Rückgang der Zahl der richtig zugeordneten Unternehmen. Trotzdem zeigt sich, dass die Verwendung traditioneller Verfahren einen großen Beitrag zur Erreichung der faktischen Anonymität auch bei wirtschaftsstatistischen Einzeldaten leisten kann. Allerdings scheint es notwendig, diese Maßnahmen mit datenverändernden Verfahren zumindest bei den großen Unternehmen zu flankieren.

Auch wenn nur ein kleiner Ausschnitt der Umsatzsteuerstatistik auf die Vertraulichkeit der Daten hin getestet wurde, erscheinen die Ergebnisse geeignet für eine Verallgemei-

13) Siehe z. B. die Ergebnisse zur Kostenstrukturerhebung im Verarbeitenden Gewerbe; siehe Vorgrimler, D.: „Reidentifikationsmöglichkeiten am Beispiel eines konkreten Angriffsszenarios“ in Gnoss, R./Ronning, G. (Hrsg.): „Anonymisierung wirtschaftsstatistischer Einzeldaten“, Band 42 der Schriftenreihe „Forum der Bundesstatistik“, Wiesbaden 2003, S. 40 ff.

14) Siehe Lenz, R.: “A graph theoretical approach to record linkage”, Joint UN-ECE/Eurostat work session on statistical data confidentiality, Luxemburg 2003.

nerung auf die gesamte Erhebung zu sein. Dabei spielen zwei Gründe eine wesentliche Rolle. Zunächst einmal sind fast alle sehr großen Unternehmen in dem betrachteten Ausschnitt enthalten und die durchschnittliche Größe der betrachteten Unternehmen liegt signifikant höher als die der nicht betrachteten (durchschnittlich 4,3 Mill. gegenüber 1 Mill. Euro Umsatz). Da sich die Annahme bestätigt hat, dass die größeren Unternehmen stärker gefährdet sind als die kleineren, ist davon auszugehen, dass die Trefferquoten sinken, sobald es gelingt, sämtliche Unternehmen in die Betrachtung mit einzubeziehen. Ebenfalls hat sich gezeigt, dass Unternehmen in geringer besetzten Wirtschaftszweigen einem erhöhten Risiko ausgesetzt sind. In dem betrachteten Ausschnitt sind nicht nur sieben der zehn am dünnsten besetzten Abteilungen enthalten, die durchschnittliche Besetzung ist mit 12 000 Unternehmen auch deutlich geringer als bei den übrigen Abteilungen, die mit durchschnittlich 94 000 Unternehmen besetzt sind. Dies wird ebenfalls dazu führen, dass die Trefferquoten bei einer Betrachtung der gesamten Umsatzsteuerstatistik geringer sein werden, als dies bei dem betrachteten Ausschnitt der Fall war. Für die Vertraulichkeit der Daten heißt das, dass man bei einer Verallgemeinerung der ermittelten Trefferquote das Risiko eher über- als unterschätzt.

Aufgrund dieser möglichen Verallgemeinerung liefern die Ergebnisse einen wichtigen Schritt hin zu einem Scientific Use File für die Umsatzsteuerstatistik. Allerdings sind die Ergebnisse für die Sicherheitsanalyse noch unvollständig, da bisher nur die Möglichkeit einer Zuordnung betrachtet wurde. Der Nutzen, den ein Angreifer durch eine Zuordnung erzielen kann, wurde bisher ausgeblendet, spielt aber zur Erreichung der faktischen Anonymität eine wesentliche Rolle. Im Beitrag „Maße für die faktische Anonymität von Mikrodaten“ in diesem Heft auf S. 621 ff. wird auf diesen Aspekt näher eingegangen. Es sind allerdings nicht nur die Analysen zur Sicherheit unvollständig; ein Scientific Use File muss darüber hinaus noch ausreichend Analysepotenzial für die Datennutzer aufweisen. Nur wenn dies gegeben ist, macht es für die statistischen Ämter Sinn, der Wissenschaft faktisch anonyme Daten als Scientific Use Files anzubieten.

Auszug aus Wirtschaft und Statistik

Herausgeber: Statistisches Bundesamt, Wiesbaden

Schriftleitung: Johann Hahlen
Präsident des Statistischen Bundesamtes
Verantwortlich für den Inhalt:
Brigitte Reimann,
65180 Wiesbaden

- Telefon: +49 (0) 6 11/75 20 86
- E-Mail: wirtschaft-und-statistik@destatis.de

Vertriebspartner: SFG Servicecenter Fachverlage
Part of the Elsevier Group
Postfach 43 43
72774 Reutlingen
Telefon: +49 (0) 70 71/93 53 50
Telefax: +49 (0) 70 71/93 53 35
E-Mail: destatis@s-f-g.com

Erscheinungsfolge: monatlich

Allgemeine Informationen über das Statistische Bundesamt und sein Datenangebot erhalten Sie:

- im Internet: www.destatis.de

oder bei unserem Informationsservice
65180 Wiesbaden

- Telefon: +49 (0) 6 11/75 24 05
- Telefax: +49 (0) 6 11/75 33 30
- E-Mail: info@destatis.de