Anonymisierungsbeschreibung

1. Einstimmung

Im Jahre 1987 wurde 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 hier, dass die Kosten einer Reidentifikation deren Nutzen übersteigen (faktische Anonymität). Dies impliziert, dass die Enthüllung von Einzelangaben in einem faktisch anonymen Datensatz nicht mit absoluter Sicherheit ausgeschlossen werden muss.

Durch die Arbeiten des Projektes „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten" kann nun eine erste faktisch anonyme Datei für die Wissenschaft (ein so genannter Scientific-Use-File), generiert aus den Daten der Kostenstrukturerhebung im Verarbeitenden Gewerbe, angeboten werden.[2] Auf dem Weg dahin war ein klassischer Zielkonflikt zu lösen: Sicherstellung der faktischen Anonymität bei gleichzeitig bestmöglichem Erhalt des Potenzials für wissenschaftliche Analysen. Die Ergebnisse haben gezeigt, dass in der Regel eine Unterdrückung oder Vergröberung von Informationen bei den qualitativen Merkmalen wie z.B. die Zusammenfassung von Wirtschaftsabteilungen oder eine Vergröberung der Regionalangabe beachtlich zur Anonymisierung beiträgt und eine vergleichsweise schwache Modifikation der im Datensatz vorhandenen quantitativen Merkmale ermöglicht. Bei den für das Scientific-Use-File auf die Originaldaten angewendeten Anonymisierungsmaßnahmen wurde daher ein großes Gewicht auf die Behandlung der qualitativen Merkmale gelegt.

2. Anonymisierungsmaßnahmen

Wie bereits erwähnt, wurde den Anregungen der Nutzer folgend ein stärkeres Gewicht auf die Behandlung der qualitativen Merkmale gelegt.

Traditionelle Anonymisierungsverfahren

In einem ersten Schritt wurde auf das ursprünglich im Merkmalskanon vorhandene Merkmal „Tätige Inhaber" verzichtet, da es sich im Laufe der Projektarbeiten als besonders reidentifikationsgefährdend und für wissenschaftliche Analysen als wenig wertvoll herausgestellt hat.

Besonders geeignet für Reidentifikationen sind regionale Angaben. Der Erhalt solcher Merkmale in einem Scientific-Use-File stellt daher für die Anonymisierung ein schwieriges Unterfangen dar. Bereits zu Beginn des Projektes „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten" wurde die Möglichkeit ausgeschlossen, einen Scientific-Use-File zu erstellen, der administrative Gebietsangaben auf der Ebene

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).

der Bundesländer oder gar einer tieferen Gliederungsebene enthält. Da aber die Auswertung nach Regionen einen wichtigen Analysebereich darstellt, wurde nach alternativen Möglichkeiten gesucht und dies vor dem Hintergrund, gleichzeitig auf datenverändernde Maßnahmen bei den quantitativen Merkmalen weitestgehend verzichten zu wollen. Als erste Möglichkeit wurde der administrative Gebietsschlüssel durch den nichtadministrativen siedlungsstrukturellen Kreistyp BBR9 und den siedlungsstrukturellen Regionstyp BBR3 ersetzt. Die sehr deutliche Verbesserung der Schutzwirkung durch diese Vergröberung der Regionalinformation wurde in (Lenz/Vorgrimler)[3] festgestellt. Allerdings sprach sich der Wissenschaftliche Begleitkreis dafür aus, anstelle dieser nicht-administrativen Schlüssel eine Ost-West-Klassifizierung einzuführen. Diese zweite Möglichkeit wurde schließlich in den Scientific-Use-File aufgenommen.

Die Daten der Kostenstrukturerhebung im Verarbeitenden Gewerbe wurden nach der Klassifikation der Wirtschaftszweige, Ausgabe 1993 (WZ93), auf der Vierstellerebene (Klasse) erhoben und aufbereitet. Diese Klassifikation ist von der europäischen Klassifikation NACE Rev.1 abgeleitet, die aufgrund der NACE-Verordnung des Rates der Europäischen Gemeinschaften seit 1995 in allen Mitgliedstaaten der Europäischen Union sowohl für die Erhebung als auch für die Darstellung der statistischen Daten anzuwenden ist.[4] Das Kodierungssystem der WZ93 unterscheidet zwischen Abschnitten (Buchstaben A-Q), Unterabschnitten (Buchstaben AA-QA), Abteilungen (Zweisteller), Gruppen (Dreisteller), Klassen (Viersteller) und Unterklassen (Fünfsteller). Der Wirtschaftsbereich „Verarbeitendes Gewerbe sowie Bergbau und Gewinnung von Steinen und Erden“ erstreckt sich über die Abschnitte C und D bzw. – in der numerischen Gliederung – über die Abteilungen 10 bis 37. Im Projekt „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten“ haben sich Datenschützer und Datennutzer darauf verständigt, bei dem hierarchischen Merkmal WZ93 die Gliederungstiefe 2 (Zweistellerebene) nicht zu unterschreiten, da hierdurch zum einen eine beachtliche Schutzwirkung und zum anderen nach Einschätzung der beteiligten Wissenschaftler für einen Scientifc-Use-File eine ausreichende Breite an Analysemöglichkeiten erhalten wird.

In den Veröffentlichungen der statistischen Ämter werden aufgrund von Geheimhaltungsaspekten die Ergebnisse einiger Wirtschaftsabteilungen nicht veröffentlicht. Es handelt sich dabei um Unternehmen der Abteilungen *10, 11, 14, 16, 23, 30, 32, 35* und *37* der WZ93. Bei den im Projekt durchgeführten Simulationen hat sich bestätigt, dass diese Abteilungen neben den Abteilungen *15, 17, 18, 19* , 22 und *34* größerer Geheimhaltung bedürfen. Um diese kritischen Abteilungen im Scientific-Use-File belassen und weitgehend auf datenverändernde Verfahren bei den quantitativen Merkmalen verzichten zu können, werden die Abteilungen *10* (Kohlenbergbau, Torfgewinnung), *11* (Gewinnung von Erdöl und Erdgas, Erbringung damit verbundener Dienstleistungen) und *14* (Gewinnung von Steinen und Erden, sonstiger Bergbau) zum Abschnitt C, die Abteilungen *15* (Ernährungsgewerbe) und *16* (Tabakverarbeitung) zum Unterabschnitt DA, die Abteilungen *17* (Textilgewerbe) und *18* (Bekleidungsgewerbe) zum Unterabschnitt

---

2 Aggregate dieser Erhebung werden in der Fachserie 4.3, „Produzierendes Gewerbe - Kostenstrukturerhebung der Unternehmen des Verarbeitenden Gewerbes sowie des Bergbaus und der Gewinnung von Steinen und Erden“, des Statistischen Bundesamtes veröffentlicht.

3 Siehe Lenz, R./Vorgrimler, D.: „Matching German Turnover Tax Statistics“, erscheint in der Reihe der Diskussionspapiere des Forschungsdatenzentrums des Statistischen Bundesamtes, Wiesbaden.

4 Für neuere Erhebungen ab dem Jahr 2003 gilt mit dem Branchenschlüssel WZ 2003 wiederum eine neue Klassifikation.

DB, die Abteilungen *21* (Papiergewerbe) und *22* (Verlags- und Druckgewerbe, Vervielfältigung) zum Unterabschnitt DE, die Abteilungen *30* (Herstellung von Büromaschinen, Dv-Geräten und -einrichtungen) und 31 (Herstellung von Geräten der Elektrizitätserzeugung, -verteilung u.ä.) zum Unterabschnitt DL sowie die Abteilungen *34* (Herstellung von Kraftwagen und Kraftwagenteilen) und *35* (Sonstiger Fahrzeugbau) zum Unterabschnitt DM zusammengefasst. Bei den Abteilungen *19* (Ledergewerbe) und *23* (Kokerei, Mineralölverarbeitung, Herstellung von Brutstoffen) wurde das Merkmal WZ93 unterdrückt. Außerdem wurde die Abteilung *37* (Recycling) aus inhaltlichen und aus Geheimhaltungsgründen herausgenommen. Obwohl die Abteilungen 32 (Rundfunk-, Fernseh- u. Nachrichtentechnik) und 33 (Medizin-, Mess-, Steuer- u. Regelungstechnik, Optik) ebenfalls zum Unterabschnitt DL zu zählen sind, werden Sie im Datensatz separat aufgeführt, da hier die Weitergabe der Zweisteller aus Sicht des Datenschutzes unbedenklich ist. Eine zusammenfassende Aufstellung der im Datensatz vorhandenen Ausprägungen des Merkmals WZ93 enthält nachfolgende Tabelle:

Wirtschaftsgliederung	WZ93-Angabe
Bergbau und Gew. v. Steinen u. Erden	C (10, 11 und 14)
Ernährungsgewerbe u. Tabakverarbeitung	DA (15 und 16)
Textil- u. Bekleidungsgewerbe	DB (17 und 18)
Holzgewerbe (oh. H. v. Möbeln)	20
Papier-, Verlags- u. Druckgewerbe	DE (21 und 22)
Chemische Industrie	DG (bzw. 24)
H. v. Gummi- u. Kunststoffwaren	DH (bzw. 25)
Glasgewerbe, Keramik, Ver. V. Steinen u. Erden	DI (bzw. 26)
Metallerzg. u. –bearbeitung	27
H. v. Metallerzeugnissen	28
Maschinenbau	DK (bzw. 29)
H. v. Büromasch., Dv-Gerät. u. einr., Gerät. d. Elektriz.erzg., -verteilung u. ä.	30 und 31
Rundfunk-, Fernseh- u. Nachrichtentechnik	32
Medizin-, Mess, Steuer- u. Regelungstechnik,Optik	33
Fahrzeugbau	DM (34 und 35)
H.v.Möbeln,Schmuck,Musikinstr., Sportger. usw	36
Sonstige: Ledergewerbe; Kokerei, Mineralölverarbeitung, H. v. Brutstoffen	19 und 23

Eindimensionale Mikroaggregation

Die im Datensatz verbleibenden 30 quantitativen Merkmale wurden eindimensional für jedes Merkmal separat mikroaggregiert.[5] Bei dieser Variante der Mikroaggregation werden zunächst die Merkmalsausprägungen je Merkmal absteigend sortiert. Dann werden tripelweise (aus den

5 Zur Methode der Mikroaggregation und anderen im Projekt untersuchten Methoden siehe Höhne, J.: „Methoden zur Anonymisierung wirtschaftsstatistischer Einzeldaten“ in Ronning, G./Gnoss, R.: „Anonymisierung wirtschaftsstatistischer Einzeldaten“, Band 42 der Schriftenreihe „Forum der Bundesstatistik“, Wiesbaden 2003, S. 69 ff.

Merkmalsausprägungen dreier benachbarter Merkmalsträger) die Durchschnittswerte ermittelt, die Originalwerte durch diese Durchschnittswerte ersetzt und wieder an die ursprüngliche Position zurücksortiert. Falls die Anzahl der Merkmalsträger nicht durch die Zahl Drei teilbar ist, so ist am Ende der absteigend sortierten Liste von Merkmalsausprägungen auch die Bildung einer Gruppe aus vier oder fünf Merkmalsträgern zulässig. Damit ist jede Merkmalsausprägung bei mindestens drei Merkmalsträgern vorhanden. Das hier skizzierte Verfahren ist für das Ziel einer möglichst vielseitigen Datennutzung, sowohl für deskriptive als auch für ökonometrische Auswertungen, das schonendste Verfahren innerhalb der Klasse der Mikroaggregationsverfahren.

3. Stellungnahme des Wissenschaftlichen Begleitkreises

In den Rückmeldungen des für das Projekt „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten“ eingerichteten Wissenschaftlichen Begleitkreises wurden die methodischen Arbeiten zur Beurteilung des Analysepotenzials, welche sowohl deskriptive Maße als auch inferenzstatistische Auswertungen in Form linearer und nichtlinearer ökonometrischer Modellierung beinhalten, als überzeugend beurteilt. Die konkrete Anonymisierungsstrategie für die Kostenstrukturerhebung im Verarbeitenden Gewerbe wurde als sorgfältig und umfassend eingestuft und der Empfehlung zur Erstellung eines Scientific-Use-Files voll zugestimmt. Insgesamt wurden die Forschungsarbeiten des Projektes als sehr erfolgreich im Hinblick auf die Untersuchungsziele bewertet.