IAW-Diskussionspapiere

IAW-Discussion Paper

9

Erste Ergebnisse von vergleichenden Untersuchungen mit anonymisierten und nicht anonymisierten Einzeldaten

am Beispiel der Kostenstrukturerhebung und der Umsatzsteuerstatistik

Martin Rosemann

Juni 2003

INSTITUT FÜR
ANGEWANDTE
WIRTSCHAFTSFORSCHUNG
Ob dem Himmelreich 1
72074 Tübingen

T: (0 70 71) 98 96-0
F: (0 70 71) 98 86-99
E-Mail: iaw@iaw.edu
Internet: www.iaw.edu

ISSN: 1617-5654

IAW-Diskussionspapiere

Das Institut für Angewandte Wirtschaftsforschung (IAW) Tübingen ist ein unabhängiges außeruniversitäres Forschungsinstitut, das am 17. Juli 1957 auf Initiative von Professor Dr. Hans Peter gegründet wurde. Es hat die Aufgabe, Forschungsergebnisse aus dem Gebiet der Wirtschafts- und Sozialwissenschaften auf Fragen der Wirtschaft anzuwenden. Die Tätigkeit des Instituts konzentriert sich auf empirische Wirtschaftsforschung und Politikberatung.

Dieses **IAW-Diskussionspapier** können Sie auch von unserer IAW-Homepage als pdf-Datei herunterladen:

http://www.iaw.edu/Publikationen/IAW-Diskussionspapiere

ISSN 1617-5654

Weitere Publikationen des IAW:

- IAW-Report (erscheinen 2-3x jährlich)
- IAW-Wohnungsmonitor Baden-Württemberg (erscheinen 4x jährlich kostenlos)
- IAW-Forschungsberichte

Möchten Sie regelmäßig eine unserer Publikationen erhalten, dann wenden Sie sich bitte an uns:

IAW Tübingen, Ob dem Himmelreich 1, 72074 Tübingen,
Telefon 07071 / 98 96-0
Fax 07071 / 98 96-99
E-Mail: iaw@iaw.edu

Aktuelle Informationen finden Sie auch im Internet unter: http://www.iaw.edu

---

Inhaltsverzeichnis

Erste Ergebnisse von vergleichenden Untersuchungen mit anonymisierten und nicht anonymisierten Einzeldaten

am Beispiel der Kostenstrukturerhebung und der Umsatzsteuerstatistik

Martin Rosemann*)

Einführung

Gegenstand des vorliegenden Beitrags ist es, erste Ergebnisse von vergleichenden Untersuchungen mit anonymisierten und nicht anonymisierten Einzeldaten aus dem Bereich der Unternehmens- und Betriebsdaten vorzustellen. Mit Hilfe dieser Ergebnisse werden erste Schlussfolgerungen gezogen, inwiefern verschiedene im Rahmen des Projekts „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten“[1)] erprobte Verfahren zur Anonymisierung von Einzeldaten das Analysepotenzial der betrachteten Datensätze verändern. Darauf aufbauend kann entschieden werden, welche Verfahren oder Verfahrensgruppen nach dem Kriterium der geringsten Verminderung des Analysepotenzials für eine tiefergehende Betrachtung im weiteren Verlauf des Projekts in Frage kommen und künftig für die faktische Anonymisierung wirtschaftsstatistischer Einzeldaten angewendet werden können.

Zunächst wird in Kapitel 1 der Versuch unternommen, zu klären, was unter dem Begriff Analysepotenzial zu verstehen ist und wie die Verringerung des Analysepotenzials operationalisiert werden kann. In Kapitel 2 wird ein Überblick über die im Rahmen dieses Beitrags betrachteten Datensätze, die Kostenstrukturerhebung und die Umsatzsteuerstatistik, gegeben[2)]. In Kapitel 3 werden die Auswirkungen traditioneller und datenverändernder Anonymisierungsverfahren auf die Umsatzsteuerstatistik einer näheren Untersuchung unterzogen. Dabei werden zunächst in Abschnitt 3.1 die angewendeten Anonymisierungsverfahren kurz erläutert. Abschnitt 3.2 widmet sich der Veränderung wichtiger Charakteristika der Verteilung. In Abschnitt 3.3 werden die Ergebnisse deskriptiver Analysen mit der Umsatzsteuerstatistik bei anonymisierten und Originaldaten verglichen. Abschnitt 3.4 zieht ein Zwischenfazit für die Umsatzsteuerstatistik. Die Untersuchung der Auswirkungen von datenverändernden Anonymisierungsverfahren auf das Analysepotenzial der Kostenstrukturerhebung wird in Kapitel 4 vorgenommen. Dabei werden in Abschnitt 4.1 die verwendeten Anonymisierungsverfahren vorgestellt. Abschnitt 4.2 untersucht die Veränderungen wichtiger Charakteristika der Verteilung. In Abschnitt 4.3 werden die Ergebnisse deskriptiver Analysen mit der Kostenstrukturerhebung bei anonymisierten und Originaldaten verglichen. In Abschnitt 4.4 wird untersucht, wie sich die Ergebnisse ökonometrischer Schätzungen bei Anwendung ver-

*) Diplom-Volkswirt Martin Rosemann, Institut für Angewandte Wirtschaftsforschung (IAW), Tübingen.

1) Das Projekt „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten“ wird von den Statistischen Ämtern getragen und vom Bundesministerium für Bildung und Forschung gefördert. Das IAW Tübingen ist an diesem Projekt als Unterauftragnehmer des Statistischen Bundesamts beteiligt.

2) Die Gründe, warum diese Datensätze im Projekt zuerst bearbeitet werden, werden ebenfalls in Kapitel 2 erläutert.

schiedener Anonymisierungsverfahren verändern und inwieweit sich die Ergebnisse für verschiedene Verfahren unterscheiden. Abschnitt 4.5 zieht ein Zwischenfazit für die Kostenstrukturerhebung. In Kapitel 5 werden die wesentlichen Ergebnisse der Untersuchungen der Veränderung des Analysepotenzials zusammengefasst, erste Schlussfolgerungen für die Verwendung von Anonymisierungsverfahren aus Sicht des Analysepotenzials gezogen und ein abschließender Ausblick auf weitere Forschungsarbeiten gegeben.

1 Operationalisierung des Analysepotenzials

Ziel ist es, Anonymisierungsverfahren danach zu beurteilen, wie stark sie das Analysepotenzial eines Datensatzes einschränken. Hierzu ist es notwendig, das Analysepotenzial zu operationalisieren und möglichst objektive Kriterien für seine Veränderung aufzustellen.

Als besonders problematisch erweist sich zunächst, dass auch das Analysepotenzial eines Originaldatensatzes nicht eindeutig und objektiv fassbar ist. Die Untersuchungsziele sind ebenso vielseitig wie die Methoden und Verfahren, mit denen sie erreicht werden sollen. Und letztlich hängt das Analysepotenzial vor allem von Art und Beschaffenheit der zugrundeliegenden Statistik ab.

Dennoch lässt sich leicht nachvollziehen, dass unabhängig von der konkreten Fragestellung bzw. dem angewendeten Verfahren bestimmte Verteilungseigenschaften für die betrachteten Merkmale auch nach der Anwendung von Anonymisierungsverfahren wenigstens annähernd erhalten bleiben sollten. Schließlich ist die gesamte in statistischen Auswertungen genutzte Information eines Datensatzes in der multivariaten Verteilung der erhobenen Variablen enthalten (vgl. Brand et al. 1999, S. 158). Erhaltenswerte Eigenschaften der multivariaten Verteilung sind insbesondere:

- Mittelwerte und Streuungsmaße der univariaten Verteilungen,
- Kovarianzen und Korrelationen zwischen den Variablen bzw. die Rangkorrelationen oder andere robuste Zusammenhangsmaße, insbesondere bei schiefen Verteilungen.

Genannt werden können auch die dritten und vierten Momente, wobei man davon ausgehen muss, dass bei der Einbeziehung von zu vielen Zielen und Kriterien Kompromisse zwischen unterschiedlichen Zielen gemacht werden müssen oder aber eine Prioritätensetzung erfolgen muss. So ist beispielsweise die Anwendung der meisten Standardverfahren, wie beispielsweise einer Regressionsschätzung, daran gebunden, dass die Mittelwerte und die Varianz-Kovarianz-Matrix im anonymisierten Datensatz näherungsweise erhalten bleiben (vgl. Brand, Bender, Kohaut 1999).

Als Kriterium für die Einschränkung des Analysepotenzials durch verschiedene Anonymisierungsverfahren dient dann die Abweichung der Kennzahl (des Charakteristikums) des anonymisierten Datensatzes von der Kennzahl des Originaldatensatzes. Angelehnt an Sebé et al. 2002 sowie Dandekar, Domingo-Ferrer und Sebé 2002 werden die mittleren Fehler

- der Mediane
- der arithmetischen Mittel
- der Varianzen
- der Kovarianzen
- der Korrelationskoeffizienten (Bravais-Pearson)
- der Rang-Korrelationskoeffzienten (Spearman)

verwendet. Diese Kriterien werden unabhängig von der betrachteten Statistik, der interessierenden Fragestellung und der anzuwenden Analysemethode zur Beurteilung der Verringerung des Analysepotenzials herangezogen.[3] Im Unterschied zu Sebé et al. 2002 sowie Dandekar, Domingo-Ferrer und Sebé 2002 werden die Maße in diesem Beitrag allerdings nicht zu einer einzigen Kennzahl verdichtet. Vielmehr werden die Kriterien einzeln betrachtet.

Folgende Kriterien werden im Rahmen dieser Arbeit untersucht (vgl. Höhne 2002):

(Bei den mit * gekennzeichneten Größen handelt es sich jeweils um die anonymisierten Werte, bei den nicht gekennzeichneten Größen um die Originalwerte. Mit d wird die Anzahl der Variablen bezeichnet.)

a) $$\frac{\sum_{j=1}^{d} \frac{\left|\overline{x_j} - \overline{x_j^*}\right|}{\left|\overline{x_j}\right|}}{d}$$ Mittlerer relativer Fehler der arithmetischen Mittel

b) $$\frac{\sum_{j=1}^{d} \sum_{1 \le i < j} \frac{\left|\mathrm{cov}_{ij} - \mathrm{cov}_{ij}^*\right|}{\left|\mathrm{cov}_{ij}\right|}}{1/2\, d(d-1)}$$ Mittlerer relativer Fehler der Kovarianzen

c) $$\frac{\sum_{j=1}^{d} \frac{\left|\mathrm{var}_{jj} - \mathrm{var}_{jj}^*\right|}{\left|\mathrm{var}_{jj}\right|}}{d}$$ Mittlerer relativer Fehler der Varianzen

d) $$\frac{\sum_{j=1}^{d} \sum_{i=1}^{d} \frac{\left|\mathrm{cov}_{ij} - \mathrm{cov}_{ij}^*\right|}{\left|\mathrm{cov}_{ij}\right|}}{1/2\, d(d+1)}$$ Mittlerer relativer Fehler der Varianz-Kovarianzmatrix

e) $$\frac{\sum_{j=1}^{d} \sum_{1 \le i < j} \left|r_{ij} - r_{ij}^*\right|}{1/2\, d(d-1)}$$ Mittlerer absoluter Fehler der Korrelationskoeffizienten

3) Wenig überzeugend ist die Betrachtung der Abweichung der Einzelwerte, wie sie in Sebé et al. 2002 sowie Dandekar, Domingo-Ferrer und Sebé 2002 ebenfalls vorgenommen wird. Zum einen ist es gerade die Kernidee datenverändernder Anonymisierungsverfahren, dass die Werte voneinander abweichen, zum anderen ist mit der Abweichung der einzelnen Werte noch keine Aussage über die Verteilungseigenschaften und damit über das Analysepotenzial verbunden.

f) $$\frac{\sum_{j=1}^{d}\sum_{1\le i<j}\left|s_{ij}-s_{ij}^{*}\right|}{\frac{1}{2}d(d-1)}$$ Mittlerer absoluter Fehler der Rangkorrelationen

Eine abschließende Bewertung über die Einschränkung des Analysepotenzials kann dennoch erst im Zusammenhang mit den Auswirkungen von Anonymisierungsmaßnahmen auf unterschiedliche Arten von Analysen vorgenommen werden. Im Weiteren werden daher verschiedene Arten von Analysen durchgeführt, um die Effekte von Anonymisierungsverfahren auf das Analysepotenzial zu untersuchen. Dabei werden zum einen beispielhafte deskriptive Auswertungen vorgenommen. Von besonderer Bedeutung ist die deskriptive Auswertung von Teilmassen, wie z.B. die Berechnung bestimmter Durchschnittsgrößen nach Wirtschaftszweigen und/oder Beschäftigtengrößenklassen sowie ihrer Rangzahlen. Zum anderen werden die Auswirkungen der Anonymisierungsverfahren auf die Koeffizienten verschiedener ökonometrischer Modelle sowie auf die entsprechenden Test-Statistiken untersucht. Vorgegangen wird also in drei Schritten:

1. Vergleich der arithmetischen Mittel, der Korrelationsstruktur, der Rangkorrelationen, der Kovarianzen und der Varianzen; Kriterium: Abweichungen dieser Größen
2. Vergleich deskriptiver Kennzahlen von Teilmassen (z.B. FuE-Intensitäten nach Wirtschaftszweigen; FuE-Intensitäten nach Größenklassen, FuE-Intensitäten nach Wirtschaftszweigen und Größenklassen); Kriterien: Veränderung der Mittelwerte, Veränderung der Rangfolgen zwischen den Teilmassen
3. Durchführung ökonometrischer Schätzungen (lineare und nichtlineare Modelle); Kriterium: Veränderung der Koeffizienten und der Teststatistiken.

Bei der Untersuchung der Veränderung von Koeffizienten im Rahmen ökonometrischer Modelle muss insbesondere analysiert werden,

- inwiefern sich die statistische Signifikanz eines Einflussfaktors verändert,
- inwiefern sich bei gegebener statistischer Signifikanz das Vorzeichen eines Koeffizienten verändert,
- ob die Veränderung der Werte von Koeffizienten statistisch signifikant ist.

Erst all dies gemeinsam liefert die Grundlage für die Entscheidung darüber, ob ein Anonymisierungsverfahren vor dem Hintergrund der Zielsetzung der weitgehenden Erhaltung des Analysepotenzials geeignet ist.

2 Zugrundeliegende Datensätze: Umsatzsteuerstatistik und Kostenstrukturerhebung

Im Rahmen des Projekts „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten“ wird für sechs verschiedene Statistiken untersucht, ob und gegebenenfalls mit welchen Verfahren sie faktisch anonymisiert werden können.

Für die ersten Untersuchungen wurden zunächst die Umsatzsteuerstatistik und die Kostenstrukturerhebung (KSE) ausgewählt, weil

- die Umsatzsteuerstatistik aufgrund einer sehr hohen Zahl an Untersuchungseinheiten (2,9 Millionen Unternehmen) sowie einer geringen Zahl an Variablen tendenziell eher leichter zu anonymisieren sein dürfte und der Einsatz traditioneller Anonymisierungsverfahren hier zunächst am ehesten ausreichend erscheint (vgl. Vorgrimler 2002).
- die Kostenstrukturhebung auf der einen Seite aufgrund einer vergleichsweise geringen Zahl von Untersuchungseinheiten (16.918), einer großen Zahl von Variablen sowie eines großen vermuteten Zusatzwissens nur schwer zu anonymisieren ist und daher vermutlich in jedem Fall der Einsatz datenverändernder Verfahren notwendig sein dürfte. Auf der anderen Seite dürfte sich die Kostenstrukturerhebung eines recht großen Interesses von Seiten der Nutzer erfreuen.

Damit werden gleich zu Beginn des Projektes zwei Datensätze bearbeitet, die sich sowohl in ihren Nutzungsmöglichkeiten als auch in ihren Charakteristika und damit in den notwendigen Anonymisierungsverfahren deutlich unterscheiden. In den Tabellen 2.1 und 2.2 sind die zur Verfügung stehenden Variablen beider Statistiken aufgeführt.

Tabelle 2.1: Variablen der Umsatzsteuerstatistik

1. Regionalbezug (BBR-Schlüssel, sog. „Neuner-Kategorie“)
2. Wirtschaftszweig (WZ93)
3. Dauer der Steuerpflicht (typisiert)
4. Organschaft nach § 2 Abs. 2 Nr. 2 UstG (0= nein, 1 = ja)
5. Rechtsform

Jahreswerte in Euro

6. Lieferungen und Leistungen
7. Steuerpflichtige Lieferungen und Leistungen
8. Zu 16 %
9. Zu 7 %
10. Steuerfreie Lieferungen und Leistungen
11. Mit Vorsteuerabzug
12. Innergemeinschaftliche Lieferungen und Leistungen
13. Ohne Vorsteuerabzug
14. Umsatzsteuer vor Abzug der Vorsteuer
15. Für Lieferungen und Leistungen
16. Für innergemeinschaftliche Erwerbe
17. Abziehbare Vorsteuer
18. Für Lieferungen und Leistungen
19. Aus Rechnungen anderer Unternehmen
20. Einfuhrumsatzsteuer
21. Für innergemeinschaftliche Erwerbe
22. Vorauszahlungssoll
23. Nachrichtlich: innergemeinschaftliche Erwerbe

Vorjahreswerte in Euro

24. Lieferungen und Leistungen
25. Vorauszahlungssoll

Tabelle 2.2: Variablen der Kostenstrukturerhebung

1. Wirtschaftszweig (WZ 93)
2. Regionalbezug (BBR-Schlüssel, sog. "Neuner-Kategorie")
3. Beschäftigtengrößenklasse
4. Tätige Inhaber
5. Angestellte und Arbeiter
6. Teilzeitbeschäftigte
7. Teilzeitbeschäftigte umgerechnet in Vollzeiteinheiten
8. Tätige Personen insgesamt
9. Umsatz aus eigenen Erzeugnissen
10. Umsatz aus Handelsware
11. Gesamtumsatz (entspricht nicht der Summe aus 9. und 10.)
12. Anfangsbestand an unfertigen und fertigen Erzeugnissen aus eigener Produktion gemessen
13. Endbestand an unfertigen und fertigen Erzeugnissen aus eigener Produktion gemessen am
14. Bestandveränderung an unfertigen/fertigen Erzeugnissen
15. Gesamtleistung/Bruttoproduktionswert
16. Anfangsbestand an Rohstoffen und sonstigen fremdbezogenen Vorprodukten, Hilfs- und
17. Endbestand an Rohstoffen und sonstigen fremdbezogenen Vorprodukten, Hilfs- und Be-
18. Verbrauch an Rohstoffen
19. Energieverbrauch
20. Anfangsbestand an Handelsware gemessen am Umsatz aus Handelsware
21. Endbestand an Handelsware gemessen am Umsatz aus Handelsware
22. Einsatz an Handelsware
23. Bruttogehalts- und -lohnsumme
24. Gesetzliche Sozialkosten
25. Sonstige Sozialkosten
26. Kosten für Leiharbeitnehmer
27. Kosten für Lohnarbeiten
28. Kosten für Reparaturen
29. Mieten und Pachten
30. Sonstige Kosten
31. Fremdkapitalzinsen
32. Kosten insgesamt
33. Bruttowertschöpfung zu Faktorkosten
34. Nettowertschöpfung zu Faktorkosten
35. Gesamtaufwendungen für innerbetriebliche Forschung und Entwicklung
36. Anzahl der für Forschung und Entwicklung eingesetzten Lohn- und Gehaltsempfänger

Die Umsatzsteuerstatistik ist eine Sekundärstatistik, die auf die Daten zurückgreift, die bei der Finanzverwaltung im Rahmen des Umsatzsteuer-Voranmeldungs- und -Vorauszahlungsverfahrens anfallen. Erfasst werden alle Unternehmen mit einem Jahresumsatz (ohne Umsatzsteuer) von über 32.500 DM (16.617 €), die Umsatzsteuer-Voranmeldungen abgeben. Seit 1996 wird die Umsatzsteuerstatistik jährlich durchgeführt[4)] (Statistisches Bundesamt 2002).

4) Für die im Rahmen des Projekts „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten“ durchgeführten Untersuchungen liegen Daten der Umsatzsteuerstatistik für das Jahr 2000 vor.

Die Kostenstrukturerhebung im Verarbeitenden Gewerbe erfasst als hochrechnungsfähige Stichprobe maximal 18.000 Unternehmen mit 20 und mehr Beschäftigten; die Befragung erfolgt zentral durch das Statistische Bundesamt im Wege der Selbstausfüllung durch die Unternehmen. Die in der Stichprobe gewonnenen Ergebnisse werden auf die Gesamtheit der Unternehmen mit 20 und mehr Beschäftigten hochgerechnet. Diese Stichprobe wird i.d.R. alle vier Jahre neu gezogen, so dass kleinere und mittlere Unternehmen durch Rotation entlastet werden können. Unternehmen mit 500 und mehr Beschäftigten, aber auch Unternehmen in Wirtschaftszweigen mit geringer Besetzungszahl, werden zur Sicherstellung der Qualität der Ergebnisse vollständig einbezogen. Den Ergebnissen für das Berichtsjahr 1999 liegt eine neue Stichprobenauswahl zugrunde. In dieser Stichprobe werden rd. 43 % der Unternehmen des Verarbeitenden Gewerbes sowie des Bergbaus und der Gewinnung von Steinen und Erden mit 20 und mehr Beschäftigten erfasst. Die Konstruktion des Stichprobenplans garantiert, dass diese Unternehmen zu 76% zur Gesamtzahl der tätigen Personen und zu 84 % zum Gesamtumsatz im Berichtskreis beitragen (Statistisches Bundesamt 2002).

3 Untersuchungen für die Umsatzsteuerstatistik

3.1 Angewendete Anonymisierungsverfahren

Für die Prüfung der Veränderung des Analysepotenzials durch Anonymisierungsmaßnahmen im Rahmen dieses Beitrags stehen sowohl mit sogenannten traditionellen Verfahren (probe)anonymisierte Datensätze als auch mit datenverändernden Verfahren bearbeitete Datensätze zur Verfügung.[5] Bei traditionellen Verfahren werden im Unterschied zu den so genannten datenverändernden Verfahren keine generellen Veränderungen der einzelnen Werte vorgenommen (vgl. Brand 2000, Höhne 2003 sowie Ronning et al. 2002).

3.1.1 Auf die Umsatzsteuerstatistik angewendete traditionelle Verfahren[6]

Anonymisiert werden nur die Überschneidungsmerkmale Umsatz (Lieferungen und Leistungen), Regionaltyp, Wirtschaftszweig und Rechtsform sowie solche Merkmale, die mit dem Umsatz hoch korreliert sind.[7]

Im Einzelnen werden folgende Maßnahmen vorgenommen:

- Die Wirtschaftszweige werden bis zu einem Umsatz von 500 Millionen Euro als Zweisteller (WZ 93) ausgewiesen, ab einem Umsatz von 500 Millionen Euro lediglich als Einsteller.
- Die Wirtschaftszweigklassifikationen werden so zusammengefasst, dass jeweils mindestens 3.500 Einheiten in einer Klasse vertreten sind.
- Die Rechtsform wird so umkodiert (vergröbert), dass nur noch vier verschiedene Kategorien ausgewiesen werden: Kategorie 1 (Personengesellschaften) umfasst

5) Die Anonymisierung wurde im Rahmen des Projekts „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten“ durch das Statistische Bundesamt und das Statistische Landesamt Berlin vorgenommen.

6) Für einen Überblick über die traditionellen Anonymisierungsverfahren siehe insbesondere Müller et al. 1991 sowie Ronning et al. 2002. Einen Überblick über die datenverändernden Verfahren geben Höhne 2003, Ronning et al. 2002, Brand 2000 sowie Gottschalk 2002.

7) Vgl. hierzu Vorgrimler 2002.

Einzelunternehmen, OHG und KG. Kategorie 2 (Kapitalgesellschaften) umfasst AG und GmbH. Kategorie 3 umfasst Erwerbs- und Wirtschaftsgenossenschaften und Körperschaften des öffentlichen Rechts. Kategorie 4 umfasst die sonstigen Rechtsformen.

- Für das Merkmal Umsatz wird für Unternehmen mit einem Umsatz von mindestens 500 Millionen Euro das Replacement-Verfahren angewendet. D.h. alle Umsätze, die zwischen 500 Millionen und 1 Mrd. Euro liegen, werden durch das arithmetische Mittel dieser Umsatzwerte ersetzt. Das gleiche wird für Umsätze über 1 Mrd. Euro durchgeführt.
- Bei Unternehmen mit einem Umsatz von weniger als 500 Mio. € werden Rundungen vorgenommen. Umsatzwerte zwischen 50 und 500 Mio. € werden auf die ersten beiden Stellen gerundet, Umsätze von weniger als 50 Mio. € auf die erste Stelle.

Außerdem werden verschieden große Stichproben gezogen. Variante 1 stellt wie der Originaldatensatz die volle Erhebung dar (Trad 1). Variante 2 ist eine 80 %-Stichprobe (Trad 1, 80 %), Variante 3 eine 50 %-Stichprobe (Trad 1, 50 %).

3.1.2 Auf die Umsatzsteuerstatistik angewendete datenverändernde Verfahren

Bei der Anonymisierung der Umsatzsteuerstatistik wird das Verfahren SAFE des Statistischen Landesamts Berlin in zwei Varianten angewendet (vgl. Evers, Höhne 1999 und Höhne 2003). Dabei wird zuerst eine Lösung gesucht, die unter ausschließlicher Betrachtung der diskreten Fälle keine Einzel- oder Zweierfälle mehr aufweist. Dieser Lösung werden die originalen Sätze zugeordnet (mit größter Ähnlichkeit und Veränderungen bei möglichst kleinen stetigen Merkmalen). Anschließend werden die stetigen Merkmale nach zwei Varianten anonymisiert.

- SAFE1: Trippelbildung innerhalb der Gruppen der diskreten Merkmale (gleiche Ausprägungen bei den diskreten Merkmalen). Innerhalb der identischen Gruppen werden die Unternehmen absteigend nach einem ausgewählten dominierenden Merkmal sortiert und anschließend für jeweils drei benachbarte Werte alle stetigen Merkmalswerte durch das arithmetische Mittel ersetzt.
- SAFE2: Anonymisierung der stetigen Werte mit eindimensionaler Mikroaggregation (für jedes Merkmal getrennt). Die Gruppengröße weist eine Mindestgröße von drei Elementen sowie eine minimale Schwankungsbreite von 7 % auf.

3.2 Vergleich wesentlicher Charakteristika der Verteilungen

Die Ergebnisse in Tabelle 3.1. zeigen, dass die beiden Varianten des SAFE zu keiner Veränderung der arithmetischen Mittel führen. Die Anwendung der traditionellen Verfahren ohne zusätzliche Stichprobenziehung führt nur zu einer relativ geringen Veränderung der arithmetischen Mittel um 0,2 %. Durch die Stichprobenziehung ergibt sich dann eine zusätzliche größere Veränderung. Bei der Veränderung der Varianzen schneidet Verfahren SAFE2 mit einer Veränderung um 8,6 % mit Abstand am besten ab. Alle anderen Verfahren haben Veränderungen von mindestens 30 % zur Folge. Bei den Kovarianzen schneiden alle Verfahren mit eine Veränderung von ca. 60 % ungefähr gleich ab. Die geringsten Veränderungen der Korrelationskoeffizienten verursacht Verfahren SAFE1. Die anderen Verfahren liegen bei der

Veränderung der Korrelationskoeffizienten etwa gleichauf. Bei den Rangkorrelationen schneidet Verfahren SAFE1 mit Abstand am schlechtesten ab. Alle anderen Verfahren haben durchschnittliche absolute Veränderungen von etwa 0,005. Bei den traditionellen Verfahren kann noch festgehalten werden, dass die Zusatzmaßnahme der Stichprobenziehung die arithmetischen Mittel deutlich stärker verändert als die Varianzen, Kovarianzen, Korrelationskoeffizienten und Rangkorrelationen.

Insgesamt lässt sich aus diesen Veränderungen der Verteilungscharakteristika noch keine eindeutige Antwort geben, wie stark welches Anonymisierungsverfahren das Analysepotenzial einschränkt.

Tabelle 3.1: Veränderung von Verteilungscharakteristika durch die Anonymisierung der Umsatzsteuerstatistik

Verfahren	Mittlerer relativer Fehler			Mittlerer absoluter Fehler	
	Arithmetische Mittel	Varianzen	Kovarianzen	Korrelationen	Rangkorrelationen
	in %	in %	in %	(x 100)	(x 100)
SAFE1	0,0	56,4	57,4	3,5	8,5
SAFE2	0,0	8,6	56,6	10,8	0,5
Trad1	0,2	33,9	60,3	15,2	0,5
Trad 1, 80 %	2,5	31,5	60,1	15,0	0,5
Trad 1, 50 %	3,5	37,1	64,9	16,2	0,5

Quelle: IAW-Berechnungen

3.3 Vergleich deskriptiver Auswertungen

Für die Umsatzsteuerstatistik werden folgende deskriptive Auswertungen vorgenommen:

- Berechnung der Anteile der steuerpflichtigen Lieferungen und Leistungen zu 7 % und zu 16 % an allen Lieferungen und Leistungen sowie an allen steuerpflichtigen Lieferungen und Leistungen nach Wirtschaftszweigen. Damit kann analysiert werden, für welche Wirtschaftszweige der ermäßigte Mehrwertsteuersatz von 7 % von größerer und für welche er von geringerer Bedeutung ist.
- Berechnung der Anteile der innergemeinschaftlichen Lieferungen und Leistungen und der sonstigen steuerfreien Lieferungen und Leistungen mit Vorsteuerabzug (Exporte außerhalb der EU) an den Lieferungen und Leistungen. Damit kann analysiert werden, welche Wirtschaftszweige stärker und welche weniger stark exportabhängig sind.
- Berechnung der Anteile der abziehbaren Vorsteuer aus Einfuhrumsatzsteuer sowie für innergemeinschaftliche Lieferungen und Leistungen an der gesamten abziehbaren Vorsteuer. Damit kann analysiert werden, welche Wirtschaftszweige stärker und welche weniger stark importabhängig sind.

In Tabelle 3.2 sind beispielhaft für die Untersuchung der Bedeutung des ermäßigten Mehrwertsteuersatzes die Veränderungen der arithmetischen Mittel nach WZ-Zweistellern (auf Basis der im Rahmen der traditionellen Anonymisierung teilweise zusammengefassten Wirtschaftszweige sowie der Ränge der Wirtschaftszweige nach arithmetischen Mitteln dargestellt, die sich durch die Anwendung der verschiedenen Anonymisierungsverfahren ergeben.

Tabelle 3.2: Veränderung der arithmetischen Mittel sowie der Ränge der Wirtschaftszweige nach arithmetischen Mitteln durch die Anonymisierung der Umsatzsteuerstatistik bei der Untersuchung der Bedeutung des ermäßigten Mehrwertsteuersatzes

Verfahren	Durchschnittliche relative Veränderung der arithmetischen Mittel in %		Durchschnittliche absolute Veränderung der Ränge	
	Umsatzanteile zu 7 % am Gesamtumsatz	Umsatzanteile zu 7 % am steuerpflichtigen Umsatz	Umsatzanteile zu 7 % am Gesamtumsatz	Umsatzanteile zu 7 % am steuerpflichtigen Umsatz
SAFE1	0,0	0,0	0,0	0,0
SAFE2	1,2	1,2	0,0	0,0
Trad 1	14,3	10,9	0,6	0,7
Trad 1, 80 %	19,8	15,2	1,0	1,0
Trad 1, 50 %	30,4	26,7	1,6	1,4

Quelle: IAW-Berechnungen

Es ist zu erkennen, dass die beiden Varianten des SAFE-Verfahrens, insbesondere SAFE1, die Ergebnisse kaum verändern und von den getesteten Verfahren am besten abschneiden. Die traditionellen Verfahren schneiden hingegen deutlich schlechter ab. Es wird deutlich, dass sowohl die zunächst durchgeführten traditionellen Verfahren als auch die anschließende Stichprobenziehung eine erhebliche Einschränkung des Analysepotenzials bedeuten.

Tabelle 3.3: Minimale und maximale Veränderung der arithmetischen Mittel sowie der Ränge der Wirtschaftszweige nach arithmetischen Mitteln durch die Anonymisierung der Umsatzsteuerstatistik für alle Auswertungen

Verfahren	Intervall der durchschnittlichen Veränderungen der arithmetischen Mittel nach WZ-Zweistellern in %	Intervall der durchschnittlichen absoluten Veränderung der Ränge der Zweisteller nach arithmetischen Mitteln
SAFE1	[0,0 ; 0,0]	[0,0 ; 0,0]
SAFE2	[0,4 ; 2,5]	[0,0 ; 3,5]
Trad 1	[0,7 ; 16,2]	[0,2 ; 2,5]
Trad 1, 80%	[1,1 ; 20,9]	[0,7 ; 2,8]
Trad 1, 50%	[1,6 ; 30,4]	[1,4 ; 3,2]

Quelle: IAW-Berechnungen

In Tabelle 3.3 sind die Ergebnisse für alle oben aufgeführten Auswertungen, die mit der Umsatzsteuerstatistik vorgenommen wurden, verdichtet dargestellt. Ausgewiesen ist jeweils die geringste und die größte Abweichung von den Originalergebnissen, die sich bei allen Auswertungen beobachten lässt. Dabei werden die für das in Tabelle 3.2 aufgeführte Beispiel festgehaltenen Ergebnisse bestätigt. Deutlich wird, dass mit SAFE1 bei den deskriptiven Auswertungen der Umsatzsteuerstatistik bessere Ergebnisse erzielt werden als mit SAFE2. Die traditionellen Verfahren schneiden zwar bei der Veränderung der arithmetischen Mittel selbst schlechter ab als das Verfahren SAFE2, nicht jedoch bei der Veränderung der Ränge.

3.4 Ein Zwischenfazit für die Umsatzsteuerstatistik

Das aussichtsreichste Verfahren bei der Anonymisierung der Umsatzsteuerstatistik scheint nach den ersten durchgeführten ersten Auswertungen das Verfahren SAFE des Statistischen Landesamts Berlin in der Variante SAFE1 zu sein. Die angewendeten traditionellen Verfahren führen zu einer größeren Einschränkung des Analysepotenzials als die SAFE-Verfahren. Dies gilt nicht für alle berechneten Charakteristika der Verteilung, allerdings werden Mittelwerte und Korrelationskoeffizienten deutlich stärker verändert. Bei der deskriptiven Analyse ist die Abweichung der Ergebnisse zwischen den beiden Verfahrenstypen bei den arithmetischen Mitteln innerhalb der Wirtschaftszweige stärker ausgeprägt als bei der Veränderung der Ränge. Dennoch ist das Ergebnis bei allen durchgeführten Auswertungen anzutreffen. Es scheint, dass die traditionellen Verfahren vor der Stichprobenziehung eine größere Einschränkung des Analysepotenzials mit sich bringen als die anschließende Stichprobenziehung.

4 Untersuchungen für die Kostenstrukturerhebung

4.1 Angewendete Anonymisierungsverfahren

Bei der Probe-Anonymisierung der Kostenstrukturerhebung werden bisher nur datenverändernde Verfahren angewendet. Es werden folgende Verfahrensgruppen[8)] zur Anonymisierung getestet und in Bezug auf die Verringerung des Analysepotenzials verglichen:

- Mikroaggregation (MA)
- SAFE (Verfahren des Statistischen Landesamtes Berlin)
- Rank-Swapping (RSWP)
- Latin Hypercube Sampling (LHS)

a) Zu den angewendeten Varianten der Mikroaggregation (MA)

Die Gruppengröße beträgt mindestens drei Einheiten. Die Originalwerte werden durch die arithmetischen Mittel der Gruppe ersetzt.

Folgende Varianten werden getestet:

- Alle stetigen Variablen werden gemeinsam betrachtet. D.h. die Gruppen werden nach der kleinsten euklidischen Distanz aus allen stetigen Merkmalen gebildet (MA1g).

8) Zu den Verfahren im Einzelnen vgl. Höhne 2003 sowie Ronning et al. 2002.

- Es wird eine Blockung in Variablen für Handelstätigkeit einerseits und andere Variablen andererseits vorgenommen. Die Gruppenbildung im Rahmen der Mikroaggregation erfolgt für die beiden Blöcke getrennt (MA2g).
- Jede der insgesamt 33 stetigen Variablen wird getrennt mikroaggregiert (MA33g).

b) Zu den angewendeten Varianten des SAFE

Beim Verfahren SAFE wird zunächst eine Lösung gesucht, die unter ausschließlicher Betrachtung der diskreten Fälle keine Einzel- oder Zweierfälle mehr aufweist. Es wird gleichzeitig gewährleistet, dass die Fehler bei allen daraus aggregierbaren Häufigkeitsverteilungen einen minimalen Maximalfehler haben und die Anzahl aller Objekte erhalten bleibt. Dieser Lösung werden die originalen Sätze zugeordnet (mit größter Ähnlichkeit und Veränderungen bei möglichst kleinen stetigen Merkmalen) (vgl. Höhne 2003).

Anschließend werden die stetigen Merkmale in zwei Varianten anonymisiert:

- SAFE1: Trippelbildung innerhalb der Gruppen der diskreten Merkmale (gleiche Ausprägungen bei den diskreten Merkmalen): Innerhalb der identischen Gruppen werden die Unternehmen absteigend nach einem ausgewählten dominierenden Merkmal (hier: tätige Personen) sortiert. Anschließend werden für jeweils drei benachbarte Werte alle stetigen Merkmalswerte durch das arithmetische Mittel ersetzt.
- SAFE2: Anonymisierung der stetigen Werte mit eindimensionaler Mikroaggregation analog zu MA33g. Die Gruppengröße weist eine Mindestgröße von drei Elementen sowie eine minimale Schwankungsbreite von 7 % auf.

Bei den Mikroaggregationsverfahren sowie bei SAFE2 werden Felder, die sich als Summe aus anderen Feldern berechnen lassen bzw. unmittelbar und logisch aus anderen Feldern hervorgehen (z.B. Beschäftigtengrößenklasse) bei der Anonymisierung zunächst vernachlässigt und im Anschluss an die Anonymisierung der anderen Felder aus diesen neu berechnet.[9)]

c) Zu den angewendeten Varianten des Rank-Swapping (RSWP)

Alle Variablen (stetige und diskrete) werden getrennt voneinander bearbeitet. Zunächst werden die Einheiten für jedes Merkmal nach der Höhe des entsprechenden Merkmalswertes sortiert. Bei jeder Variable wird ein Tausch von Merkmalswerten zwischen jeweils zwei Merkmalsträgern in einem vorab festgelegten Nachbarschaftsbereich vorgenommen. Die Nachbarschaftsbereiche betragen 10 % (RSWP 10p), 5 % (RSWP 5p) und 1 % (RSWP 1p) der Anzahl aller Untersuchungseinheiten. (Zum Beispiel würden bei n = 1000 Beobachtungen und einem Nachbarschaftsbereich von 1 % innerhalb der nächstgelegenen 10 Beobachtungen getauscht).

d) Zur angewendeten Variante des Latin Hypercube Sampling (LHS)

Latin Hypercube Sampling (LHS) ist ein Simulationsverfahren, das von R. Dandekar für die Anonymisierung von Einzeldaten vorgeschlagen wurde (Dandekar 1993, Dandekar, Domingo-Ferrer, Sebé 2002, Dandekar, Cohen, Kirkendall 2002). Das Verfahren erlaubt es, (beliebige) univariate Verteilungen zu simulieren. Darüber hinaus werden die Rangkorrelationen erhalten (vgl. Höhne 2003 und Ronning et al. 2002). Für die hier vorgenommenen Untersuchungen werden die Unternehmen der KSE in vier Gruppen unterteilt. Es wird danach unter-

9) So ergibt sich die neue Beschäftigtengrößenklasse aus dem anonymisierten Wert für die tätigen Personen insgesamt.

schieden, ob Handel[10] und/oder Forschung und Entwicklung betrieben wird[11]. Für diese vier Gruppen werden die stetigen Variablen mit LHS anonymisiert. Aus den anonymisierten Teildatensätzen werden anschließend mithilfe einer stark vereinfachten Variante der in Dandekar, Domingo-Ferrer, Sebé 2002 vorgeschlagenen Methode LHS Hybrid Daten gebildet, indem die anonymisierten Teildatensätze anhand der Ausprägung eines einzigen Merkmals, nämlich der Bruttogehalts- und Lohnsumme, den Ursprungsdatensätzen und damit den diskreten Variablen zugeordnet werden.[12] An dieser Stelle muss deutlich darauf hingewiesen werden, dass die nach diesem Verfahren (im folgenden als LHS1 bezeichnet) gebildeten LHS Hybrid Daten eigentlich keine Auswertungen nach Ausprägungen der diskreten Variablen (sog. Teilmassenbetrachtungen) zulassen. Dies machen die in 4.3 und 4.4 zusammengestellten Untersuchungsergebnisse deutlich, bei denen die Daten ausschließlich auf Teilmassenebene, bzw. unter Einbeziehung der Ausprägungen der diskreten Variablen ausgewertet werden.[13]

Anzumerken ist noch, dass bei der Anwendung von LHS1 die Zahl Merkmalsträger erhöht werden kann. Schließlich werden die Werte der stetigen Variablen synthetisch erzeugt. Bei der durchgeführten Variante des LHS wird die Zahl der Unternehmen von 16.918 auf 17.100 erhöht. Es kommt also vor, dass einer originalen Kombination von diskreten Merkmalen zwei unterschiedliche synthetische Sätze an stetigen Merkmalswerten zugeordnet werden.

4.2 Vergleich wesentlicher Charakteristika der Verteilungen

Tabelle 4.1 zeigt die Veränderung der wesentlichen Verteilungscharakteristika durch die Anonymisierungsverfahren. Es ist zu erkennen, dass durch das Rank-Swapping-Verfahren die ersten und zweiten Momente nicht verändert werden, dies ist verfahrensimmanent. Allerdings werden die Varianz-Kovarianzmatrix und die Korrelationskoeffizienten deutlich verändert. Ein ähnliches Bild ergibt sich für das Simulationsverfahren LHS1. Hier werden die Momente der univariaten Verteilung kaum verändert[14], die Kovarianzen und die Korrelationsstruktur hingegen deutlich. Die Mikroaggregationsverfahren MA1g und MA2g sowie das Verfahren SAFE1 führen zu den höchsten Fehlern bei den Varianzen. Dies liegt daran, dass durch die Durchschnittsbildung innerhalb der Gruppen bei Mikroaggregation und SAFE automatisch ein Teil der Variation verloren geht. Bei den Fehlern der Kovarianzen liegen sie im mittleren Bereich, die Korrelationskoeffizienten verringern sich nur geringfügig.[15] Die ge-

10) Entscheidendes Kriterium hierfür ist, dass die Variable „Einsatz an Handelsware“ größer Null ist.

11) Herangezogen wird die Variable „Anzahl der für Forschung und Entwicklung eingesetzten Lohn- und Gehaltsempfänger“.

12) Dandekar, Domingo-Ferrer, Sebé 2002 schlagen vor, diese Paarbildung nicht anhand eines, sondern anhand aller im LHS Datensatz vorhandenen Merkmale vorzunehmen.

13) Falls es zu den Vorgaben für das Anonymisierungsverfahren gehört, dass solche Teilmassenbetrachtungen angestellt werden können sollen, wird in Dandekar, Cohen, Kirkendall 2002 vorgeschlagen, das LHS Verfahren getrennt auf diese Teilmassen anzuwenden und, wo dies wegen zu geringer Gruppenbesetzungen nicht möglich ist, LHS auf einen Datensatz anzuwenden, der die betreffenden kategorialen Merkmale bereits enthält. Allerdings ist es nur schwer vorstellbar, für unterschiedliche Teilmassenauswertungen unterschiedlich anonymisierte Scientific use files zur Verfügung zu stellen. Dennoch sind Verbesserungen gegenüber der Variante LHS1 in dieser Hinsicht möglich.

14) Die Veränderungen entstehen lediglich durch die zusätzlich geschaffenen künstlichen Merkmalsträger.

15) Die Abweichungen der arithmetischen Mittel bei den Mikroaggregationsverfahren ergeben sich dadurch, dass die Lagerbestände im Ursprungsdatensatz als Anteile an Umsatzgrößen ausgewiesen sind. Vor der Anonymisierung wird auf die absoluten Werte zurückgerechnet. Diese werden anonymisiert und anschließend neu auf die ebenfalls anonymisierten Umsatzwerte bezogen.

ringsten Abweichungen bei den Kovarianzen und Korrelationskoeffizienten weist das Mikroaggregationsverfahren MA33g auf. Auch bei den Varianzen und arithmetischen Mitteln ist die Abweichung gering. Das Verfahren SAFE2 schließlich weist ebenso wie MA33g geringe Abweichungen bei den Korrelationskoeffizienten, arithmetischen Mitteln und Varianzen auf, allerdings eine vergleichsweise hohe Abweichung der Kovarianzen.

Es zeigt sich, dass die Rank-Swapping-Verfahren und das Latin Hypercube Sampling zwar die univariaten Verteilungen erhalten, die Zusammenhänge zwischen den Variablen aber nur sehr unzureichend abbilden, während die Mikroaggregationsverfahren und SAFE zwar zu stärkeren Veränderungen bei den univariaten Verteilungen führen, die Zusammenhänge aber weniger zerstören.

Bei den Rangkorrelationen ergibt sich nochmals ein völlig verändertes Bild. Hier weist nach dem Verfahren MA33g das Verfahren RSWP 1p die geringste Veränderung auf. Es folgen RSWP 5p, SAFE2 und LHS1. Die größten Abweichungen bei den Rangkorrelationen weisen MA1g und MA2g auf. Ebenso wie bei der Umsatzsteuerstatistik verursacht auch bei der KSE das Verfahren SAFE1 eine vergleichsweise hohe Abweichung bei den Rangkorrelationen.

Diese unterschiedlichen Ergebnisse machen nochmals deutlich, dass es sinnvoll ist, die Maßzahlen, die sich aus der Veränderung von Verteilungscharakteristika ergeben, nicht zu einer einzigen Kennzahl zu verdichten, sondern die Auswirkung der Verfahren auf konkrete Analysen zu untersuchen, um ein geschlossenes Gesamtbild zu erhalten.

Tabelle 4.1: Veränderung von Verteilungscharakteristika durch die Anonymisierung der Kostenstrukturerhebung

Verfahren	Mittlerer relativer Fehler			Mittlerer absoluter Fehler	
	Arithmetisches Mittel	Varianzen	Varianz-Kovarianzmatrix	Korrelationen	Rangkorrelationen
	in %	in %	in %[16)]	(x 100)	(x 100)
MA1g	3,5	21,3	75,8	5,8	9,0
MA2g	2,5	23,4	62,0	4,8	6,8
MA33g	0,0	5,9	21,2	2,4	0,0
SAFE1	2,8	46,9	88,6	4,4	6,6
SAFE2	0,0	7,3	96,4	3,8	0,5
RSWP	0,0	0,0	131,9	35,4	1,6
RSWP 5p	0,0	0,0	130,8	34,5	0,5
RSWP 1p	0,0	0,0	147,6	31,6	0,1
LHS1	1,0	0,6	219,6	36,2	0,8

Quelle: Berechnungen des Statistischen Landesamts Berlin und des IAW

16) Im Unterschied zu den Abschnitten 1.2 und 3.2 wird hier die Abweichung der gesamten Varianz-Kovarianzmatrix berechnet.

4.3 Vergleich deskriptiver Auswertungen

Im Rahmen deskriptiver Auswertungen mit der Kostenstrukturerhebung werden die FuE-Beschäftigungs- und Ausgabenintensitäten nach Wirtschaftszweigen und Beschäftigtengrößenklassen untersucht.[17)] Betrachtet werden jeweils die Veränderung der Kennzahlen selbst sowie die Veränderungen in der Rangstruktur.[18)]

Die FuE-Beschäftigungsintensitäten werden als Anteil der für Forschung und Entwicklung eingesetzten Beschäftigten an der Anzahl der insgesamt im Unternehmen tätigen Personen bestimmt. Die FuE-Ausgabenintensitäten werden als Anteil der Ausgaben für Forschung und Entwicklung am Gesamtumsatz berechnet.

Die durchschnittlichen Intensitäten werden jeweils für WZ-Viersteller, WZ-Zweisteller und Beschäftigtengrößenklassen sowie für die möglichen Kombinationen aus WZ-Vier- und Zweistellern und den verschiedenen Beschäftigtengrößenklassen berechnet.

Beispielhaft ist in den Tabellen 4.2 bis 4.4 zunächst dargestellt, wie sich die arithmetischen Mittel bzw. die Ränge durch die Anonymisierungsverfahren bei einer Untersuchung der FuE-Intensitäten aller Unternehmen nach WZ-Zweistellern, WZ-Vierstellern sowie WZ-Zweistellern und Beschäftigtengrößenklassen verändern.

Tabelle 4.2: Veränderung der arithmetischen Mittel und der Ränge bei der Untersuchung der FuE-Beschäftigungs- und FuE-Ausgabenintensitäten nach WZ-Zweistellern

Verfahren	Durchschnittliche relative Veränderung der arithmetischen Mittel in %		Durchschnittliche absolute Veränderung der Ränge	
	FuE-Beschäftigungs-intensitäten	FuE-Ausgaben-intensitäten	FuE-Beschäftigungs-intensitäten	FuE-Ausgaben-intensitäten
MA1g	117,4	82,3	1,7	1,9
MA2g	129,9	107,6	0,6	1,1
MA33g	0,2	0,2	0,0	0,0
SAFE1	18,9	21,0	0,8	0,9
SAFE2	12,3	15,6	0,9	0,5
RSWP 10p	747,6	974,4	5,2	4,6
RSWP 5p	310,7	478,6	3,1	4,2
RSWP 1p	97,1	96,4	2,2	2,5
LHS1	70,0	156,6	2,2	2,8

Quelle: IAW-Berechnungen

17) Vergleichbare Untersuchungen wurden auch für die Kostenstruktur (Anteile verschiedener Kostenpositionen am Gesamtumsatz) der Unternehmen durchgeführt. Die Auswirkungen der Anonymisierungsmaßnahmen sind allerdings ähnlich.

18) Die Berechnungen wurden am Institut für Angewandte Wirtschaftsforschung in Tübingen durchgeführt. Vorarbeiten und wertvolle Hinweise stammen von Professor Dr. Joachim Wagner (Universität Lüneburg).

Es ist zu erkennen, dass das Verfahren der Mikroaggregation, bei dem alle Variablen getrennt voneinander bearbeitet werden (MA33g), mit Abstand am besten abschneidet. Gefolgt wird es von den beiden Varianten des SAFE, wobei SAFE2 in der Regel besser abschneidet als SAFE1. Die einzige Ausnahme stellt die Untersuchung nach WZ-Vierstellern dar. Die schlechtesten Ergebnisse liefern die Rank-Swapping-Verfahren, insbesondere RSWP 10p und RSWP 5p. MA2g, MA1g und die Hybrid Daten (LHS1) liegen im Mittelfeld. Es zeigt sich also, dass die Tatsache, dass die Rankswapping-Verfahren die univariaten Verteilungen der Ursprungsvariablen erhalten, selbst bei einfachen deskriptiven Auswertungen nicht ausreicht, um ähnliche Ergebnisse wie mit den Originaldaten zu erreichen. Dies liegt daran, dass bei der Berechnung der Intensitäten der Quotient aus zwei Variablen gebildet wird, die bei der Anwendung des Rank-Swappings völlig unabhängig voneinander getauscht werden. Außerdem werden beim Rank-Swapping zusätzlich auch die Ausprägungen der diskreten Variablen getauscht.

In den Tabellen 4.5 und 4.6 wird jeweils die Bandbreite der Abweichungen, die sich durch die einzelnen Anonymisierungsverfahren ergeben, ausgewiesen. Dabei wird danach differenziert, ob es sich um zwei- oder dreidimensionale Auswertungen[19)] handelt. Die Abweichungen der Ränge werden dadurch normiert, dass durch die Gesamtzahl der Ränge geteilt wird.

Im Wesentlichen bestätigt auch die Betrachtung der Bandbreiten die obigen Ergebnisse der obigen Beispiele. MA33g zeigt sich als das überlegene Verfahren gefolgt von SAFE2. Im Durchschnitt führen LHS1 und die Rank-Swapping-Verfahren zu den größten Veränderungen. Zu beachten ist allerdings, dass dies für die Veränderung der Mediane und seiner Ränge nicht uneingeschränkt gilt.

Tabelle 4.3: Veränderung der arithmetischen Mittel und der Ränge bei der Untersuchung der FuE-Beschäftigungs- und FuE-Ausgabenintensitäten nach WZ-Vierstellern

Verfahren	Durchschnittliche relative Veränderung der arithmetischen Mittel in %		Durchschnittliche absolute Veränderung der Ränge	
	FuE-Beschäftigungs-intensitäten	FuE-Ausgaben-intensitäten	FuE-Beschäftigungs-intensitäten	FuE-Ausgaben-intensitäten
MA1g	203,4	222,2	32,7	30,0
MA2g	269,6	282,6	21,4	24,0
MA33g	0,2	0,2	0,5	0,3
SAFE1	56,0	87,0	21,2	21,8
SAFE2	119,4	138,8	24,2	25,8
RSWP 10p	7037,2	6348,8	69,1	66,7
RSWP 5p	953,8	1100,9	56,2	56,2
RSWP 1p	285,6	354,6	46,9	43,2
LHS1	222,9	597,6	32,9	39,0

Quelle: IAW-Berechnungen

19) Zweidimensional: Nach WZ-Zweistellern, WZ-Vierstellern oder Beschäftigtengrößenklassen. Dreidimensional: Nach WZ-Zweistellern oder WZ-Vierstellern in Verbindung mit Beschäftigtengrößenklassen.

Tabelle 4.4: Veränderung der arithmetischen Mittel und der Ränge bei der Untersuchung der FuE-Beschäftigungs- und Ausgabenintensitäten nach WZ-Zweistellern und Beschäftigtengrößenklassen

Verfahren	Durchschnittliche relative Veränderung der arithmetischen Mittel in %		Durchschnittliche absolute Veränderung der Ränge	
	FuE-Beschäftigungs-intensitäten	FuE-Ausgaben-intensitäten	FuE-Beschäftigungs-intensitäten	FuE-Ausgaben-intensitäten
MA1g	89,7	87,7	13,5	13,9
MA2g	95,7	143,0	12,8	12,8
MA33g	0,2	0,2	0,7	0,7
SAFE1	58,1	82,7	8,8	8,7
SAFE2	10,4	9,3	8,5	8,0
RSWP 10p	934,7	929,1	35,1	32,4
RSWP 5p	583,0	678,7	25,9	23,7
RSWP 1p	69,7	103,8	13,1	12,8
LHS1	225,6	407,4	17,4	18,9

Quelle: IAW-Berechnungen

Tabelle 4.5: Bandbreiten der Veränderungen der arithmetischen Mittel sowie der Mediane der FuE-Beschäftigungs- und Ausgabenintensitäten und ihrer Ränge bei zweidimensionalen Analysen (Wirtschaftszweige <u>oder</u> Beschäftigtengrößenklassen)

Verfahren	Intervall der durchschnittlichen Veränderungen der arithm. Mittel		Intervall der durchschnittlichen Veränderungen der Mediane	
	Veränderung der Werte in %	Veränderung der Ränge (normiert)	Veränderung der Werte in %	Veränderung der Ränge (normiert)
MA1g	[3,6 ; 41,9]	[0,0 ; 0,1]	[3,1 ; 61,1]	[0,0 ; 0,2]
MA2g	[1,9 ; 25,2]	[0,0 ; 0,1]	[1,0 ; 36,0]	[0,0 ; 0,1]
MA33g	[0,0 ; 0,2]	[0,0 ; 0,1]	[0,0 ; 0,4]	[0,0 ; 0,1]
RSWP 10p	[34,3 ; 123,0]	[0,3 ; 0,5]	[3,2 ; 86,1]	[0,0 ; 0,4]
RSWP 5p	[10,7 ; 92,5]	[0,0 ; 0,5]	[2,1 ; 83,4]	[0,1 ; 0,3]
RSWP 1p	[5,9 ; 96,9]	[0,1 ; 0,3]	[0,9 ; 73,5]	[0,1 ; 0,3]
SAFE1	[3,9 ; 126,3]	[0,0 ; 0,1]	[3,9 ; 42,5]	[0,0 ; 0,2]
SAFE2	[0,1 ; 21,5]	[0,0 ; 0,1]	[0,1 ; 42,2]	[0,0 ; 0,1]
LHS1	[10,4 ; 126,3]	[0,0 ; 0,4]	[4,5 ; 75,0]	[0,0 ; 0,4]

Quelle: IAW-Berechnungen

Tabelle 4.6: Bandbreiten der Veränderungen der arithmetischen Mittel sowie der Mediane der FuE-Beschäftigungs- und Ausgabenintensitäten und ihrer Ränge bei dreidimensionalen Analysen (Wirtschaftszweige und Beschäftigtengrößenklassen)

Verfahren	Intervall der durchschnittlichen Veränderungen der arithm. Mittel		Intervall der durchschnittlichen Veränderungen der Mediane	
	Veränderung der Werte (in %)	Veränderung der Ränge (normiert)	Veränderung der Werte (in %)	Veränderung der Ränge (normiert)
MA1g	[6,8 ; 570,6]	[0,1 ; 0,1]	[5,9 ; 798,7]	[0,1 ; 0,2]
MA2g	[5,5 ; 305,1]	[0,1 ; 0,1]	[4,9 ; 401,6]	[0,1 ; 0,1]
MA33g	[0,1 ; 0,3]	[0,0 ; 0,0]	[0,1 ; 0,5]	[0,0 ; 0,0]
RSWP 10p	[91,7 ; 747,3]	[0,3 ; 0,3]	[15,8 ; 803,5]	[0,2 ; 0,3]
RSWP 5p	[42,8 ; 1184,9]	[0,2 ; 0,3]	[15,1 ; 1368,1]	[0,2 ; 0,3]
RSWP 1p	[20,1 ; 606,0]	[0,1 ; 0,2]	[10,1 ; 451,6]	[0,1 ; 0,2]
SAFE1	[7,0 ; 309,8]	[0,1 ; 0,1]	[6,9 ; 423,3]	[0,1 ; 0,2]
SAFE2	[2,4 ; 94,2]	[0,0 ; 0,1]	[2,5 ; 162,9]	[0,0 ; 0,1]
LHS1	[19,4 ; 2225,2]	[0,2 ; 0,4]	[14,3 ; 2292,5]	[0,3 ; 0,3]

Quelle: IAW-Berechnungen

4.4 Vergleich ökonometrischer Schätzungen

Um die Auswirkungen der verschiedenen Anonymisierungsverfahren auf ökonometrische Modelle zu untersuchen, gibt es grundsätzlich zwei Vorgehensweisen. Zum einen können theoretische Überlegungen darüber angestellt werden, wie sich Anonymisierungsmaßnahmen auf die Eigenschaften (Erwartungstreue, Konsistenz, Effizienz) von Schätzern auswirken (vgl. hierzu Lechner und Pohlmeier 2003), zum anderen können die Auswirkungen empirisch überprüft werden, indem verschiedene Modelle sowohl für die Originaldaten als auch für die anonymisierten Daten geschätzt werden und die Veränderung der Koeffizienten und der statistischen Signifikanz von Einflussfaktoren untersucht wird. Solche Simulationen sind insbesondere deshalb notwendig, weil es sich bei den Anonymisierungsverfahren teilweise um sehr komplizierte Algorithmen handelt, deren Auswirkungen nicht ohne weiteres theoretisch abgeleitet werden können[20]. Sinnvoll ist dabei insbesondere, sowohl lineare als auch nichtlineare Regressionsmodelle (insbesondere Probit-, Logit- und Tobitmodelle) mit in die Untersuchungen einzubeziehen. Dies kann im Rahmen dieses Beitrags nur für eine sehr eingeschränkte Fragestellung vorgenommen werden.

4.4.1 Die untersuchten Modelle: OLS-, Probit- und Tobitmodelle zur Erklärung der Forschungs- und Entwicklungsintensitäten

Erklärt werden soll die Höhe der FuE-Beschäftigungs- und Ausgabenintensitäten wie sie in Abschnitt 4.3 definiert wurden. Den Regressionsmodellen liegt kein theoretisches Modell zugrunde, vielmehr orientieren sie sich am vorhandenen Datenbestand. Als Einflussgrößen werden die Wirtschaftszweige auf Zweistellerebene (WZ93), die Beschäftigten (in Tausend), die quadrierten Beschäftigten (in Millionen), die Nettowertschöpfung, der Energieverbrauch,

20) Die theoretische Ableitung der Auswirkungen der Anonymisierungsmaßnahmen auf die Eigenschaften der Schätzer könnte aber auch ein Kriterium zur Auswahl von Verfahren bei der Erstellung von Scientific-use-files sein (Lechner und Pohlmeier 2003).

sowie die Anteile der Vorleistungen, Personalausgaben und Fremdkapitalzinsen am Gesamtumsatz[21] verwendet.

Für die Erklärung der Höhe der FuE-Intensitäten werden zunächst normale OLS-Regressionen geschätzt. Aufgrund der extrem linkssteilen Verteilung der Intensitäten werden alternativ auch die logarithmierten Intensitäten als abhängige Variable verwendet. Allerdings gehen durch die Logarithmierung im Originaldatensatz etwa dreiviertel aller Beobachtungen verloren, da sie eine FuE-Intensität von Null aufweisen. Zur Erklärung des qualitativen Unterschieds zwischen vorhandener FuE-Tätigkeit und nicht vorhandener FuE-Tätigkeit werden Probit-Modelle geschätzt.[22] Um den qualitativen Unterschied zwischen einer vorhandenen FuE-Tätigkeit und einer nicht vorhandenen FuE-Tätigkeit einerseits und die Höhe der FuE-Intensitäten in Abhängigkeit von den verschiedenen Einflussvariablen andererseits erklären zu können, werden ergänzend noch Tobit-Modelle angewendet.[23]

4.4.2 Veränderung der Ergebnisse der Schätzungen durch die angewendeten Anonymisierungsverfahren

Beispielhaft sind zunächst in Tabelle 4.9 die Ergebnisse der OLS-Regression der logarithmierten FuE-Beschäftigungsintensitäten auf die verschiedenen Einflussfaktoren abgebildet.

Tabelle 4.9 a): Ergebnisse von OLS-Schätzungen für die logarithmierten FuE-Beschäftigungsintensitäten (Mikroaggregation und SAFE), P-Werte in Klammern

	(0) Original	(1) MA1g	(2) MA2g	(3) MA33g	(4) SAFE1	(5) SAFE2
Energieverbrauch (Mio)	-0.001	-0.003	-0.004	-0.002	-0.003	-0.002
	(0.107)	(0.000)***	(0.000)***	(0.012)**	(0.001)***	(0.011)**
Vorleistungsquote	0.000	0.001	0.000	0.000	-0.001	0.000
	(0.862)	(0.356)	(0.926)	(0.850)	(0.262)	(0.641)
Personalausgabenquote	-0.001	-0.005	0.004	-0.001	-0.009	-0.001
	(0.282)	(0.002)***	(0.016)**	(0.311)	(0.000)***	(0.347)
Zinsaufwandsquote	0.053	0.086	0.075	0.053	0.059	0.052
	(0.000)***	(0.000)***	(0.000)***	(0.000)***	(0.000)***	(0.000)***
WZ93 Nr. 11[24]	1.625	-0.369	-0.449	1.785	4.044	2.571
	(0.087)*	(0.469)	(0.383)	(0.058)*	(0.000)***	(0.004)***
WZ93 Nr. 14	1.929	-0.312	0.015	2.149	3.995	2.323
	(0.015)**	(0.399)	(0.965)	(0.006)***	(0.000)***	(0.003)***
WZ93 Nr. 15	1.628	-0.274	-0.007	1.843	3.407	1.835
	(0.032)**	(0.374)	(0.981)	(0.014)**	(0.000)***	(0.015)**
WZ93 Nr. 16	1.989	-0.673	0.077	2.151	5.085	2.149
	(0.028)**	(0.113)	(0.869)	(0.016)**	(0.000)***	(0.013)**
WZ93 Nr. 17	2.301	0.429	0.272	2.519	4.144	2.485
	(0.002)***	(0.168)	(0.385)	(0.001)***	(0.000)***	(0.001)***
WZ93 Nr. 18	3.164	0.597	0.265	3.382	4.744	3.240
	(0.000)***	(0.062)*	(0.408)	(0.000)***	(0.000)***	(0.000)***
WZ93 Nr. 19	2.123	0.227	0.086	2.341	3.532	2.309
	(0.007)***	(0.501)	(0.799)	(0.003)***	(0.000)***	(0.003)***

21) Die letzten vier Variablen sollen insbesondere die Struktur des Unternehmens abbilden.

22) Hierfür wird eine abhängige Variable gebildet, die bei positiver FuE-Intensität den Wert 1 annimmt, bei einer FuE-Intensität von Null den Wert Null.

23) Zu den mikroökonometrischen Probit- und Tobit-Modellen vgl. insbesondere Ronning 1991.

24) Statistisches Bundesamt: Klassifikation der Wirtschaftszweige, Ausgabe 1993 (WZ93)

	(0) Original	(1) MA1g	(2) MA2g	(3) MA33g	(4) SAFE1	(5) SAFE2
WZ93 Nr. 20	1.794	-0.066	0.017	2.013	3.711	1.979
	(0.019)**	(0.837)	(0.958)	(0.008)***	(0.000)***	(0.009)***
WZ93 Nr. 21	1.598	-0.083	-0.103	1.816	3.542	1.789
	(0.036)**	(0.792)	(0.743)	(0.016)**	(0.000)***	(0.018)**
WZ93 Nr. 22	2.002	0.151	0.011	2.215	3.792	2.161
	(0.010)**	(0.631)	(0.971)	(0.004)***	(0.000)***	(0.005)***
WZ93 Nr. 23	2.764	1.111	0.594	2.997	4.125	3.005
	(0.000)***	(0.001)***	(0.097)*	(0.000)***	(0.000)***	(0.000)***
WZ93 Nr. 24	2.949	1.078	0.835	3.163	5.065	3.128
	(0.000)***	(0.000)***	(0.007)***	(0.000)***	(0.000)***	(0.000)***
WZ93 Nr. 25	2.111	0.608	0.215	2.325	4.262	2.290
	(0.005)***	(0.048)**	(0.488)	(0.002)***	(0.000)***	(0.002)***
WZ93 Nr. 26	2.042	0.344	0.178	2.259	4.043	2.225
	(0.007)***	(0.264)	(0.567)	(0.002)***	(0.000)***	(0.003)***
WZ93 Nr. 27	1.541	0.217	0.067	1.764	3.524	1.735
	(0.042)**	(0.484)	(0.829)	(0.019)**	(0.000)***	(0.021)**
WZ93 Nr. 28	2.067	0.629	0.243	2.281	4.132	2.250
	(0.006)***	(0.041)**	(0.431)	(0.002)***	(0.000)***	(0.003)***
WZ93 Nr. 29	2.644	1.071	0.727	2.856	4.917	2.824
	(0.000)***	(0.000)***	(0.018)**	(0.000)***	(0.000)***	(0.000)***
WZ93 Nr. 30	3.784	1.713	1.527	3.989	5.988	3.931
	(0.000)***	(0.000)***	(0.000)***	(0.000)***	(0.000)***	(0.000)***
WZ93 Nr. 31	2.847	1.313	0.878	3.060	5.017	3.031
	(0.000)***	(0.000)***	(0.005)***	(0.000)***	(0.000)***	(0.000)***
WZ93 Nr. 32	3.322	1.625	1.255	3.530	5.593	3.512
	(0.000)***	(0.000)***	(0.000)***	(0.000)***	(0.000)***	(0.000)***
WZ93 Nr. 33	3.215	1.576	1.243	3.429	5.535	3.385
	(0.000)***	(0.000)***	(0.000)***	(0.000)***	(0.000)***	(0.000)***
WZ93 Nr. 34	2.527	1.007	0.594	2.735	4.621	2.709
	(0.001)***	(0.001)***	(0.057)*	(0.000)***	(0.000)***	(0.000)***
WZ93 Nr. 35	2.400	1.030	0.696	2.595	4.850	2.550
	(0.002)***	(0.001)***	(0.030)**	(0.001)***	(0.000)***	(0.001)***
WZ93 Nr. 36	2.287	0.425	0.295	2.504	4.144	2.473
	(0.002)***	(0.169)	(0.342)	(0.001)***	(0.000)***	(0.001)***
WZ93 Nr. 37	2.613	0.454	0.435	2.832	3.950	2.803
	(0.011)**	(0.223)	(0.283)	(0.006)***	(0.000)***	(0.006)***
Beschäftigte (Tausend)	-0.002	-0.105	-0.093	-0.010	0.279	-0.010
	(0.916)	(0.000)***	(0.002)***	(0.614)	(0.000)***	(0.605)
Quadratbesch. (Mio)	-0.000	-0.001	-0.001	-0.000	-0.002	-0.000
	(0.041)**	(0.000)***	(0.000)***	(0.001)***	(0.000)***	(0.001)***
Nettowertsch. (Mio)	0.000	0.002	0.002	0.001	-0.001	0.001
	(0.026)**	(0.000)***	(0.000)***	(0.002)***	(0.006)***	(0.002)***
Konstante	-1.259	-0.186	-0.178	-1.482	-3.693	-1.459
	(0.096)*	(0.574)	(0.594)	(0.048)**	(0.000)***	(0.052)*
Beobachtungen	4518	8572	8082	4518	8893	4518
Bestimmtheitsmaß	0.213	0.219	0.149	0.214	0.287	0.210
Korr. Bestimmtheitsmaß	0.207	0.217	0.145	0.209	0.284	0.204
F-Wert	37.952	75.038	43.941	38.208	111.248	37.274

* signifikant zu 10%; ** signifikant zu 5%; *** signifikant zu 1%

Quelle: IAW-Berechnungen

Tabelle 4.9 b): Ergebnisse von OLS-Schätzungen für die logarithmierten FuE-Beschäftigungsintensitäten (Rank-Swapping und LHS1), P-Werte in Klammern

	(0) Original	(6) RSWP 10p	(7) RSWP 5p	(8) RSWP 1 p	(9) LHS1
Energieverbrauch (Mio)	-0.001	-0.003	-0.002	-0.001	-0.004
	(0.107)	(0.003)***	(0.157)	(0.170)	(0.000)***
Vorleistungsquote	0.000	-0.000	0.000	0.000	0.000
	(0.862)	(0.416)	(0.312)	(0.352)	(0.743)
Personalausgabenquote	-0.001	0.000	0.000	-0.000	-0.000
	(0.282)	(0.831)	(0.602)	(0.719)	(0.923)
Zinsaufwandsquote	0.053	-0.001	0.002	0.026	0.012
	(0.000)***	(0.409)	(0.277)	(0.000)***	(0.019)**
WZ93 Nr. 11	1.625	-1.589	0.000	-0.799	-0.131
	(0.087)*	(0.278)	(.)	(0.513)	(0.905)
WZ93 Nr. 14	1.929	0.710	-0.218	-0.444	1.440
	(0.015)**	(0.520)	(0.773)	(0.512)	(0.111)
WZ93 Nr. 15	1.628	0.350	-0.150	-0.707	1.172
	(0.032)**	(0.737)	(0.812)	(0.250)	(0.172)
WZ93 Nr. 16	1.989	-0.061	-0.469	-0.391	0.899
	(0.028)**	(0.958)	(0.578)	(0.749)	(0.389)
WZ93 Nr. 17	2.301	0.162	0.143	-0.182	1.175
	(0.002)***	(0.877)	(0.822)	(0.768)	(0.171)
WZ93 Nr. 18	3.164	0.302	0.205	0.526	1.346
	(0.000)***	(0.775)	(0.752)	(0.407)	(0.128)
WZ93 Nr. 19	2.123	0.134	0.478	-0.538	0.939
	(0.007)***	(0.902)	(0.496)	(0.419)	(0.295)
WZ93 Nr. 20	1.794	0.406	-0.444	-0.637	0.984
	(0.019)**	(0.700)	(0.496)	(0.310)	(0.259)
WZ93 Nr. 21	1.598	0.510	-0.215	-0.770	0.896
	(0.036)**	(0.628)	(0.744)	(0.220)	(0.301)
WZ93 Nr. 22	2.002	0.740	0.377	0.067	1.078
	(0.010)**	(0.477)	(0.553)	(0.916)	(0.225)
WZ93 Nr. 23	2.764	-0.046	0.393	0.470	1.376
	(0.000)***	(0.968)	(0.597)	(0.478)	(0.122)
WZ93 Nr. 24	2.949	0.700	0.689	0.540	1.132
	(0.000)***	(0.501)	(0.275)	(0.378)	(0.185)
WZ93 Nr. 25	2.111	0.599	0.363	-0.123	1.223
	(0.005)***	(0.565)	(0.566)	(0.841)	(0.154)
WZ93 Nr. 26	2.042	0.459	0.002	-0.375	1.163
	(0.007)***	(0.659)	(0.998)	(0.541)	(0.174)
WZ93 Nr. 27	1.541	0.427	0.112	-0.755	1.154
	(0.042)**	(0.683)	(0.861)	(0.221)	(0.179)
WZ93 Nr. 28	2.067	0.535	0.197	-0.272	1.105
	(0.006)***	(0.606)	(0.755)	(0.658)	(0.197)
WZ93 Nr. 29	2.644	0.801	0.659	0.318	1.207
	(0.000)***	(0.440)	(0.295)	(0.603)	(0.158)
WZ93 Nr. 30	3.784	1.195	0.697	0.744	1.289
	(0.000)***	(0.257)	(0.281)	(0.232)	(0.135)
WZ93 Nr. 31	2.847	1.034	1.024	0.526	1.257
	(0.000)***	(0.320)	(0.105)	(0.391)	(0.142)
WZ93 Nr. 32	3.322	1.010	1.135	0.991	1.223
	(0.000)***	(0.334)	(0.075)*	(0.108)	(0.154)
WZ93 Nr. 33	3.215	0.989	1.093	0.862	1.195
	(0.000)***	(0.342)	(0.084)*	(0.160)	(0.163)
WZ93 Nr. 34	2.527	0.841	0.681	0.146	1.075
	(0.001)***	(0.420)	(0.284)	(0.812)	(0.210)
WZ93 Nr. 35	2.400	1.150	0.528	0.194	1.368
	(0.002)***	(0.274)	(0.415)	(0.756)	(0.113)

	(0) Original	(6) RSWP 10p	(7) RSWP 5p	(8) RSWP 1 p	(9) LHS1
WZ93 Nr. 36	2.287	0.743	0.320	-0.076	1.172
	(0.002)***	(0.475)	(0.614)	(0.902)	(0.172)
WZ93 Nr. 37	2.613	0.890	-0.300	0.296	1.170
	(0.011)**	(0.402)	(0.676)	(0.667)	(0.331)
Beschäftigte (Tausend)	-0.002	-0.285	-0.204	-0.050	-0.111
	(0.916)	(0.000)***	(0.000)***	(0.000)***	(0.000)***
Quadratbesch. (Mio)	-0.000	0.002	0.001	0.000	0.001
	(0.041)**	(0.000)***	(0.000)***	(0.025)**	(0.000)***
Nettowertsch. (Mio)	0.000	-0.000	0.000	0.000	0.000
	(0.026)**	(0.651)	(0.379)	(0.299)	(0.000)***
Konstante	-1.259	0.802	0.917	1.147	0.179
	(0.096)*	(0.439)	(0.145)	(0.061)*	(0.834)
Beobachtungen	4518	4518	4518	4518	4600
Bestimmtheitsmaß	0.213	0.092	0.137	0.180	0.060
Korr. Bestimmtheitsmaß	0.207	0.085	0.131	0.174	0.053
F-Wert	37.952	14.117	22.948	30.716	9.093

* signifikant zu 10%; ** signifikant zu 5%; *** signifikant zu 1%

Quelle: IAW-Berechnungen

Es ist zu erkennen, dass insbesondere die mit den Verfahren MA33g und SAFE2 bearbeiteten Daten recht ähnliche Ergebnisse wie die Originaldaten erzeugen. Die stärksten Abweichungen ergeben sich bei den mit Rank-Swapping bearbeiteten Daten und bei den mit LHS1 erzeugten Hybrid Daten. Große Abweichungen sind aber auch bei den Verfahren MA1g und MA2g zu beobachten. Das Verfahren SAFE1 verursacht zwar weniger Veränderungen bei der statistischen Signifikanz der Einflussfaktoren, führt aber zu recht großen Veränderungen bei den Werten der Regressionskoeffizienten. Auffällig ist, dass die Verfahren MA1g und SAFE1 zu einer Erhöhung des korrigierten Bestimmtheitsmaßes und des F-Wertes führen, der Aussagen über den gemeinsamen Einfluss aller Einflussfaktoren des Modells macht. Dies ist voraussichtlich auf die bei diesen Anonymisierungsverfahren vergleichsweise starke „Glättung“ durch die Durchschnittsbildung zurückzuführen. Zu erkennen ist auch, dass sich bei den Verfahren MA1g, MA2g und SAFE1 für die vorgenommenen Schätzungen die Zahl der Beobachtungen gegenüber den Originaldaten deutlich erhöht. Dies liegt daran, dass aufgrund der vorgenommenen Durchschnittsbildung im Rahmen der Anonymisierung die Zahl der Unternehmen mit einer FuE-Beschäftigungsintensität von Null systematisch abnimmt. Dieses Phänomen ist insbesondere auch bei den vorgenommenen Probit- und Tobit-Schätzungen von Bedeutung.

Systematische Unterschiede zwischen den unterschiedlichen Schätzmodellen können dennoch bei den hier geschätzten Modellen nicht beobachtet werden. Deshalb ist in Tabelle 4.10 zusammenfassend dargestellt, wie sich die verschiedenen Anonymisierungsverfahren auf die Ergebnisse der Modellschätzungen auswirken. Als Kriterien für die Veränderung der Ergebnisse werden dabei folgende Maße verwendet:

- Anteil der Einflussfaktoren, bei denen eine Veränderung der statistischen Signifikanz beobachtet werden kann, an allen Einflussfaktoren. Als Grenzen für eine Veränderung werden Signifikanzniveaus von 1 %, 5 % und 10 % herangezogen.
- Anteil derjenigen Einflussfaktoren, die im Originaldatensatz statistisch signifikant sind, im anonymisierten Datensatz hingegen nicht (mindestens 10 % Signifikanzniveau).
- Anteil derjenigen Einflussfaktoren, die im Originaldatensatz nicht statistisch signifikant, im anonymisierten Datensatz hingegen statistisch signifikant sind (mindestens 10 % Signifikanzniveau).

- Anteil derjenigen Koeffizienten, die aufgrund der Anonymisierung ihr Vorzeichen verändern.
- Anteil derjenigen Koeffizienten, die bei gegebener statistischer Signifikanz im *Originaldatensatz* (mindestens 10 % Signifikanzniveau) aufgrund der vorgenommenen Anonymisierung ihr Vorzeichen verändern.
- Anteil derjenigen Koeffizienten, die bei gegebener statistischer Signifikanz im *anonymisierten Datensatz* (mindestens 10 % Signifikanzniveau) aufgrund der Anonymisierung ihr Vorzeichen verändern.
- Anteil derjenigen Koeffizienten, deren Werte durch die Anonymisierung statistisch signifikant verändert werden. Eine statistisch signifikante Veränderung der Werte wird dann angenommen, wenn die Koeffizientenwerte bei einem anonymisierten Datensatz außerhalb des 95 % Konfidenzintervalls des originalen Koeffizienten liegen.
- Anteil derjenigen Koeffizienten, deren Werte durch die Anonymisierung statistisch signifikant verändert werden - bei gegebener statistischer Signifikanz im *Originaldatensatz*.
- Anteil derjenigen Koeffizienten, deren Werte durch die Anonymisierung statistisch signifikant verändert werden - bei gegebener statistischer Signifikanz im *anonymisierten Datensatz*.

Auch die komprimierte Darstellung bestätigt die oben bereits erläuterten Resultate. Die besten Ergebnisse nach fast allen Kriterien zeigen die Verfahren MA33g und SAFE2. Hier ergeben sich die geringsten Veränderungen der Ergebnisse. Alle anderen Verfahren verursachen zumindest in der hier erprobten Variante recht eindeutige Veränderungen sowohl bei der statistischen Signifikanz von Einflussfaktoren als auch bei den Werten der Koeffizienten 25)26).

Tabelle 4.10: Auswirkungen von Anonymisierungsverfahren auf die Ergebnisse aller durchgeführten Modellschätzungen (OLS, Probit, Tobit)

Verfahren	(1) MA1g	(2) MA2g	(3) MA33g	(4) SAFE1	(5) SAFE2	(6) RSWP	(7) RSWP	(8) RSWP 1	(9) LHS1
Anteile jeweils in % an allen Einflussfaktoren									
Veränderung der Signifikanz von Einflussfaktoren	47	47	17	49	31	70	69	62	60
Im Original signifikant, bei Anonymisierung nicht	25	28	0	2	1	51	55	48	35
Bei Anonymisierung signifikant, im Original nicht	3	3	3	14	5	3	1	2	3

25) Der geringe Anteil derjenigen Einflussfaktoren bei den Rank-Swapping-Verfahren, der im anonymisierten Datensatz statistisch signifikant ist, im Originaldatensatz hingegen nicht, ist darauf zurückzuführen, dass hier insgesamt deutlich weniger Einflussgrößen nach der Anonymisierung überhaupt noch statistisch signifikant sind.

26) Die Modelle sind noch verbesserungsfähig. Eine optimierte Modellspezifikation wird sich vermutlich auch positiv auf die Wirkung der Anonymisierungsverfahren auswirken.

Veränderung der Vorzeichen von Koeffizienten	9	14	0	3	0	13	15	22	7
Veränderung von Vorzeichen bei gegebener Signifikanz im Original	6	9	0	1	0	9	12	19	5
Veränderung von Vorzeichen bei gegebener Signifikanz bei Anonymisierung	0	1	0	3	0	3	2	3	2
Signifikante Veränderung der Werte von Koeffizienten	65	67	6	48	11	71	75	66	31
Signifikante Veränderung der Werte von Koeffizienten bei gegebener Signifikanz im Original	62	67	3	41	10	65	72	64	29
Signifikante Veränderung der Werte von Koeffizienten bei gegebener Signifikanz bei Anonymisierung	42	47	6	11	10	30	29	25	20

(Grau unterlegt sind die nach dem jeweiligen Kriterium besten beiden Verfahren.)

Quelle: IAW-Berechnungen

4.5 Zwischenfazit für die Kostenstrukturerhebung

Diejenigen Verfahren, welche die univariaten Verteilungen ganz oder weitgehend erhalten (LHS1, Rank-Swapping) führen auf der anderen Seite zu einer weitgehenden Zerstörung der Zusammenhänge zwischen den Variablen. Dies wird insbesondere an der Veränderung der Korrelationskoeffizienten deutlich, ist aber bei den Rangkorrelationen deutlich weniger ausgeprägt. Dennoch führt dies sowohl bei den hier durchgeführten deskriptiven Analysen als auch bei den ökonometrischen Schätzungen zu deutlichen Veränderungen der Ergebnisse. Insgesamt schneidet das Verfahren am besten ab, bei dem eine getrennte Mikroaggregation für alle 33 stetigen Variablen durchgeführt wird und die diskreten Variablen unbehandelt bleiben. Die Behandlung der diskreten Variablen in Verbindung mit dieser Form der Mikroaggregation (SAFE2) liefert die zweitbesten Ergebnisse. Insbesondere bei den deskriptiven Auswertungen wird aber deutlich, dass die Behandlung der diskreten Merkmale zu einer zusätzlichen Einschränkung des Analysepotenzials führt.

5 Fazit und Ausblick

Mit diesem Beitrag wird ein erstes Zwischenfazit gezogen, wie sich verschiedene Anonymisierungsverfahren auf die Veränderung des Analysepotenzials auswirken. Beispielhaft wird dies für verschiedene Anonymisierungsverfahren mit den Daten der Umsatzsteuerstatistik und der Kostenstrukturerhebung durchgeführt. In einem ersten Schritt wird versucht, das Analysepotenzial zu operationalisieren. Es zeigt sich, dass dies nur bedingt möglich ist. Zur Bewertung der Veränderung des Analysepotenzials durch eine Anonymisierungsmaßnahme muss vielmehr sowohl die Veränderung wesentlicher Charakteristika der Verteilungen als auch die Veränderung der Ergebnisse ganz konkreter deskriptiver und ökonometrischer Analysen herangezogen werden. Neben den in diesem Beitrag untersuchten Veränderungen der arithmetischen Mittel, der Varianzen, der Kovarianzen, der Korrelationskoeffizienten und der Rangkorrelationen sollten insbesondere die Veränderungen der Mediane Gegenstand weiterer Untersuchungen sein. Sinnvoll erscheint allerdings auch die Analyse der Veränderung verschiedener Konzentrationsmaße. Bei der Untersuchung der Veränderung der Korrelationsstruktur sollte neben einer Untersuchung der Korrelationskoeffizienten auch untersucht werden, inwiefern die Zusammenhänge vor und nach der Anonymisierung statistisch signifikant sind.

Auch die vorgenommenen deskriptiven und ökonometrischen Untersuchungen mit beiden Datensätzen sind nicht ausreichend und nur beispielhaft zu verstehen. Deshalb sind weitere Untersuchungen notwendig, um die Stabilität der gefundenen Ergebnisse inhaltlich abzusichern. Für die Kostenstrukturerhebung ist daran gedacht, in einem nächsten Schritt verschiedene Produktionsfunktionen zu schätzen. Daneben sollen die hier vorgestellten ökonometrischen Modelle weiter optimiert werden.

Dennoch lassen sich aus den ersten Untersuchungen bereits Schlussfolgerungen ziehen, welche Anonymisierungsverfahren in welcher Weise das Analysepotenzial verringern. So ist deutlich geworden, dass die Verfahrensgruppe SAFE für die Umsatzsteuerstatistik im Vergleich zu den traditionellen Verfahren die besseren Ergebnisse erzielt. Dennoch sind hier weitere Verfahren zu untersuchen und zusätzliche Auswertungen vorzunehmen.

Für die Kostenstrukturerhebung kann gezeigt werden, dass diejenigen Verfahren, welche die univariaten Verteilungen annähernd erhalten, wie Rank-Swapping und Latin Hypercube Sampling zu einer teilweise sehr starken Zerstörung der Zusammenhänge zwischen den Variablen führen. Dies bedeutet insbesondere für ökonometrische Schätzungen, aber auch für die deskriptiven Auswertungen eine große Verringerung des Analysepotenzials. Die geringste Verringerung des Analysepotenzials ergibt sich durch die Mikroaggregation für jedes Merkmal getrennt. Die zusätzliche Behandlung der diskreten Variablen im Rahmen des Verfahrens SAFE2 führt zu einer zusätzlichen Einschränkung des Analysepotenzials, dennoch schneidet auch dieses Verfahren recht gut ab.

Die Verfahren unterscheiden sich ferner in ihrer Flexibilität für die Nutzer. Während bei der Anonymisierung mit Latin Hypercube Sampling (LHS) mögliche Teilmassenuntersuchungen bereits bei der Anonymisierung berücksichtigt werden müssen, ist dies bei den anderen Verfahren nicht zwingend notwendig. Allerdings kann auch die Gruppenbildung bei der Mikroaggregation analog zu SAFE1 auf gleiche Ausprägungskombinationen bei den diskreten Variablen beschränkt werden. Teilmassenuntersuchungen werden aber insbesondere für die deskriptiven Auswertungen eine wesentliche Rolle spielen.

Die Untersuchung der beiden sehr verschiedenen Datensätze Umsatzsteuerstatistik und Kostenstrukturerhebung zeigt auch, dass die Wirkungsweise der verschiedenen Anonymisie-

rungsverfahren von der Struktur und Beschaffenheit der Daten abhängig ist. Besonders schön beobachten kann man dies daran, dass das Verfahren SAFE2 bei der Umsatzsteuerstatistik zu einer stärkeren Verringerung des Analysepotenzials führt als SAFE1, während es sich bei der Kostenstrukturerhebung genau andersherum verhält. Der Grund hierfür besteht darin, dass die Umsatzsteuerstatistik aufgrund ihrer großen Zahl an Unternehmen sehr viel dichter besetzt ist und daher die bei SAFE1 durchgeführte Mikroaggregation für alle stetigen Variablen gemeinsam zu geringeren Veränderungen führt als bei der KSE.

Mit diesen ersten Untersuchungen wird deutlich, dass weitere Varianten der in diesem Beitrag betrachteten Verfahren getestet werden müssen. Auch andere Anonymisierungsverfahren, wie beispielsweise die Überlagerung mit Zufallszahlen, sollten in die Analysen einbezogen werden. Gleichzeitig wird eine zentrale Aufgabe darin liegen, die Verringerung des Analysepotenzials weiter zu systematisieren und die Zusammenhänge zwischen der Veränderung wesentlicher Verteilungscharakteristika und den Veränderungen der Ergebnisse konkreter Analysen weiter herauszuarbeiten. Das Ziel einer möglichst weitgehenden Erhaltung des Analysepotenzials muss mit der Sicherstellung der faktischen Anonymität in Einklang gebracht werden. Vor dem Hintergrund dieser beiden Ziele müssen die Anonymisierungsverfahren dann verfeinert und für die verschiedenen Erhebungen optimiert werden.

Literaturhinweise

Brand, R. (2000): Anonymität von Betriebsdaten – Verfahren zur Erfassung und Maßnahmen zur Verringerung des Reidentifikationsrisikos. In: Beiträge zur Arbeitsmarkt- und Berufsforschung 237, 2000.

Brand, R., S. Bender und S. Kohaut (1999): Möglichkeiten der Erstellung eines Scientific-Use-Files aus dem IAB-Betriebspanel. In: Spektrum Bundesstatistik, Band 14, 1999, Statistisches Bundesamt.

Dandekar, R.A. (1993): Performance improvement of Restricted Pairing Algorithm for Latin Hypercube Sampling, ASA Summer Conference, unpublished manuscript.

Dandekar, R.A., M. Cohen und N. Kirkendall (2002): Sensitive Micro Data Protection Using Latin Hypercube Sampling Technique. In: Domingo-Ferrer, Josep (Ed): Inference Control in Statistical Data Bases – From Theory to Practice. Springer, 2002.

Dandekar, R.A., J. Domingo-Ferrer und F. Sebé (2002): LHS-Based Hybrid Microdata vs Rank Swapping and Microaggregation for Numeric Microdata Protection. In: Domingo-Ferrer, Josep (Ed): Inference Control in Statistical Data Bases – From Theory to Practice. Springer, 2002.

Evers, K. und J. Höhne (1999): SAFE – Ein Verfahren zur Anonymisierung und statistischen Geheimhaltung wirtschaftsstatistischer Einzeldaten. In: Spektrum Bundesstatistik, Band 14, Wiesbaden 1999, S. 136-147.

Gottschalk, S. (2002): Anonymisierung von Unternehmensdaten. Ein Überblick und beispielhafte Darstellung anhand des Mannheimer Innovationspanels. ZEW-Discussion-Paper No. 02-23.

Höhne, J. (2002): Messung der Qualität einer anonymen Datei. Arbeitspapier der Projektgruppe „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten“.

Höhne, J. (2003): Methoden zur Anonymisierung wirtschaftsstatistischer Einzeldaten. Erscheint in: Gnoss, R. und G. Ronning (Hrsg.): Anonymisierung wirtschaftsstatistischer Einzeldaten, Forum der Bundesstatistik, Band 40, Wiesbaden.

Lechner, S. und W. Pohlmeier (2003): Schätzung ökonometrischer Modelle auf der Grundlage anonymisierter Daten. Erscheint in: Gnoss, R. und G. Ronning (Hrsg.): Anonymisierung wirtschaftsstatistischer Einzeldaten, Forum der Bundesstatistik, Band 40, Wiesbaden.

Müller, W., U. Blien, P. Knoche, H. Wirth, u.a. (1991): Die faktische Anonymität von Mikrodaten. In: Statistisches Bundesamt (Hrsg.): Forum der Bundesstatistik, Band 19, 1991.

Ronning, G. (1991): Mikroökonometrie, Berlin: Springer.

Ronning, G., R. Brand, J. Höhne, M. Rosemann und R. Wiegert (2002): Anonymisierungsverfahren – Überblick und erste Bewertung. Arbeitspapier der Projektgruppe „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten“.

Sebé, F., J. Domingo-Ferrer, J.M. Mateo-Sanz und V. Torra (2002): Post-Masking Optimization of the Tradeoff between Information Loss and Disclosure Risk in masked Microdata Sets. In: Domingo-Ferrer, Josep (Ed): Inference Control in Statistical Data Bases – From Theory to Practice. Springer, 2002.

Statistisches Bundesamt (2002): Kurzbeschreibungen der Projektdatensätze. Wiesbaden 2002.

Vorgrimler, D. (2002): Probe-Anonymisierung der Umsatzsteuerstatistik. Arbeitspapier der Projektgruppe „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten“.

IAW-Diskussionspapiere

Bisher erschienen:

Nr. 1
Das Einstiegsgeld – eine zielgruppenorientierte negative Einkommensteuer: Konzeption, Umsetzung und eine erste Zwischenbilanz nach 15 Monaten in Baden-Württemberg
Sabine Dann / Andrea Kirchmann / Alexander Spermann / Jürgen Volkert

Nr. 3
Gut betreut in den Arbeitsmarkt? Eine mikroökonomische Evaluation der Mannheimer Arbeitsvermittlungsagentur
Jürgen Jerger / Christian Pohnke / Alexander Spermann

Nr. 4
Das IAW-Einkommenspanel und das Mikrosimulationsmodell SIMST
Peter Gottfried / Hannes Schellhorn

Nr. 5
A Microeconometric Characterisation of Household Consumption Using Quantile Regression
Niels Schulze / Gerd Ronning

Nr. 6
Determinanten des Überlebens von Neugründungen in der baden-württembergischen Industrie – eine empirische Survivalanalyse mit amtlichen Betriebsdaten
Harald Strotmann

Nr. 7
Die Baulandausweisungsumlage als ökonomisches Steuerungsinstrument einer nachhaltigkeitsorientierten Flächenpolitik
Raimund Krumm

Nr. 8
Making Work Pay: U.S. American Models for a German Context?
Laura Chadwick, Jürgen Volkert

Nr. 9
Erste Ergebnisse von vergleichenden Untersuchungen mit anonymisierten und nicht anonymisierten Einzeldaten am Beispiel der Kostenstrukturerhebung und der Umsatzsteuerstatistik
Martin Rosemann