Dr. Martin Rosemann

Auswirkungen von stochastischer Überlagerung und Mikroaggregation auf die Schätzung linearer und nichtlinearer Modelle

Im November vergangenen Jahres konnte das Statistische Bundesamt im Rahmen des Gerhard-Fürst-Preises insgesamt drei hervorragende Arbeiten mit einem engen Bezug zur amtlichen Statistik auszeichnen. Die von Herrn Professor Dr. Hans Wolfgang Brachinger (Universität Freiburg Schweiz/ Université de Fribourg Suisse), dem Vorsitzenden des unabhängigen Gutachtergremiums, vorgetragenen Laudationes wurden in Ausgabe 12/2006 (S. 1229 ff.) dieser Zeitschrift bereits veröffentlicht. In der Ausgabe 03/2007 dieser Zeitschrift hat Alexander Vogel seine mit dem Gerhard-Fürst-Preis 2006 prämierte Diplomarbeit in einem eigenen Beitrag näher vorgestellt. Die Reihe mit Beiträgen über die im Jahr 2006 ausgezeichneten Arbeiten wird mit dem hier vorliegenden Beitrag von Dr. Martin Rosemann fortgesetzt. Entstanden ist seine mit einem Förderpreis in der Kategorie „Dissertationen" prämierte Arbeit „Auswirkungen datenverändernder Anonymisierungsverfahren auf die Analyse von Mikrodaten" bei Professor Dr. Gerd Ronning an der Eberhard-Karls-Universität Tübingen.

1 Einleitung

Empirische Wirtschaftsforschung und Politikberatung, aber auch die eher theoretisch orientierte ökonomische Forschung sind zur Überprüfung von Thesen und Theorien auf ein umfangreiches Datenangebot angewiesen. Dabei ist gerade die Bedeutung von Mikrodaten in den letzten Jahren gewachsen. Wichtigste Datenhalter in der Bundesrepublik Deutschland sind das Statistische Bundesamt, die Statistischen Ämter der Länder und die Bundesagentur für Arbeit mit dem ihr angegliederten Institut für Arbeitsmarkt- und Berufsforschung (IAB) sowie die Rentenversicherungsträger und die Deutsche Bundesbank. Daneben verfügen auch wirtschafts- und sozialwissenschaftliche Forschungsinstitute über nicht unwesentliche Datenbestände.

Der Zugang der Wissenschaft zu Mikrodaten der amtlichen Statistik kann lediglich im Rahmen der geltenden Rechtsvorschriften hinsichtlich Datenschutz und Geheimhaltung erfolgen.[1]) Geltende Rechtsvorschrift ist insbesondere das Gesetz über die Statistik für Bundeszwecke (Bundesstatistikgesetz – BStatG) vom 22. Januar 1987 (BGBl. I S. 462, 565). Seit dem Inkrafttreten dieses Gesetzes können Einrichtungen mit der Aufgabe unabhängiger wissenschaftlicher Forschung neben absolut anonymisierten Daten auch Einzelangaben erhalten, wenn sie nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft zugeordnet werden können (§ 16 Abs. 1 Nr. 4 und Abs. 6 BStatG). Diese Neuregelung knüpft an den Begriff der „faktischen Anonymität" an, wie er durch die European Science Foundation definiert wurde.[2]) Solcherlei faktisch anonymisierte Datenbestände werden wegen ihrer ausschließlichen Verfügbarkeit für die Wissenschaft auch als Scientific-Use-Files bezeichnet, in Abgrenzung zu den für alle Nutzer frei zugänglichen Public-Use-Files.

Für Haushalts- und Personendaten der amtlichen Statistik in Deutschland wurde bereits Anfang der 1990er-Jahre

1) Siehe Sturm, R.: „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten" in Allgemeines Statistisches Archiv, Bd. 86, 2002, S. 468 ff.

2) Siehe Gnoss, R.: „Möglichkeiten und Grenzen der Bereitstellung wirtschaftsstatistischer Einzeldaten der amtlichen Statistik für die Wissenschaft" in Statistisches Bundesamt (Hrsg.): „Möglichkeiten einer wissenschaftlichen Nutzung von Unternehmensdaten aus der amtlichen Statistik", Band 14 der Schriftenreihe „Spektrum Bundesstatistik", Wiesbaden 1999, S. 18 ff.

eine Operationalisierung der faktischen Anonymität vorgenommen[3]), in deren Folge der Wissenschaft faktisch anonymisierte Scientific-Use-Files verschiedener Haushalts- und Personenerhebungen zur Verfügung gestellt wurden. Dabei wird die faktische Anonymisierung in der Regel sichergestellt, ohne die Einzelangaben zu verfremden.

Demgegenüber wurde die Erstellung von Scientific-Use-Files im Bereich der Unternehmens- und Betriebsdaten lange Zeit für nicht realisierbar angesehen, da die Sicherstellung der faktischen Anonymität nur mit sehr starken Anonymisierungsmaßnahmen für möglich erachtet wurde. Infolgedessen wurden weitgehende Einschnitte in die Aussagekraft der Daten befürchtet.[4]) Diese Einschätzung hat verschiedene Gründe. Zunächst sind bei Unternehmens- und Betriebsdaten im Allgemeinen die Grundgesamtheiten kleiner als bei Haushalts- und Personendaten. Dies hat zur Folge, dass die Besetzungszahlen innerhalb einzelner Gruppen häufig sehr klein sind. Bei Unternehmens- und Betriebsdaten existieren dadurch mehr einzigartige, häufig auch sehr leicht zu identifizierende Fälle.[5]) Letzteres ist vor allem darauf zurückzuführen, dass die Verteilungen der quantitativen Variablen wesentlich heterogener sind.[6]) Eng damit verbunden ist die Tatsache, dass bei Unternehmens- und Betriebsdaten im Gegensatz zu Haushalts- und Personendaten häufig Dominanzen auftreten, also beispielsweise auf ein oder wenige Unternehmen ein Großteil des Umsatzes einer Branche entfällt. Ein weiteres Problem besteht darin, dass bei Unternehmens- und Betriebsdaten in der Regel größere Stichprobenauswahlsätze anzutreffen sind. Es treten, zumindest in bestimmten Größenklassen, sogar Vollerhebungen auf. Zuletzt unterscheiden sich die Daten von Unternehmen und Betrieben sehr stark in ihrer Größe. Insbesondere gibt es nur sehr wenige große Einheiten.[7])

All diese Faktoren machen es potenziellen Datenangreifern leichter, Unternehmen in formal anonymisierten Datenbeständen zu erkennen, als dies bei Personen- und Haushaltsdaten der Fall ist. Will ein Angreifer jedoch ein, mehrere oder viele Unternehmen erkennen, so benötigt er entsprechendes Zusatzwissen, das ihm eine Zuordnung ermöglicht. Beispielsweise benötigt er Kenntnisse über die Branchenzugehörigkeit, die Rechtsform, den Standort, den Umsatz und die Beschäftigtenzahl der gesuchten Unternehmen. Dabei ist leicht einsichtig, dass solches Zusatzwissen für Unternehmen und Betriebe in weitaus größerem Umfang, deutlich leichter zugänglich und besser aufbereitet zur Verfügung steht als für Haushalte und Personen. Dies ergibt sich insbesondere aus Publizitätspflichten von Unternehmen, der Existenz von allgemeinen Unternehmensdatenbanken und Bilanzdatenbanken.[8]) Es kommt hinzu, dass Unternehmen ab einer gewissen Größe oft zu mehreren Erhebungen meldepflichtig sind.[9]) Damit ergeben sich Probleme insbesondere für die Anonymisierung größerer Unternehmen, die häufiger innerhalb bestimmter Merkmalskombinationen einzigartig sind und für die aufgrund von Publizitätspflichten vergleichsweise viel Zusatzwissen zur Verfügung steht.[10])

Allerdings erhielt das Ziel, auch Scientific-Use-Files für wirtschaftsstatistische Einzeldaten zu erstellen, durch den wachsenden und zunehmend deutlicher artikulierten Bedarf der Wissenschaft nach Einzeldaten für Unternehmen und Betriebe einen neuen Impuls. So hat der Statistische Beirat im Jahr 1996 in seinen Vorschlägen für ein Rahmenkonzept zur „Neuordnung der amtlichen Statistik“ eine Verbesserung des Zugangs zu anonymisierten Mikrodaten angemahnt, um die Belastung der Befragten durch zusätzliche Erhebungen zu vermeiden.[11]) Defizite in der Verfügbarkeit wirtschaftsstatistischer Einzeldaten wurden insbesondere auch im Gutachten der Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik (KVI) problematisiert.[12])

Mit dem Projekt „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten“ wurden die Möglichkeiten der faktischen Anonymisierung von Unternehmens- und Betriebsdaten ausführlich untersucht und die Grundlagen für die Erstellung von Scientific-Use-Files auch in diesem Bereich gelegt.[13]) Erste Scientific-Use-Files im Bereich der Unternehmens- und Betriebsdaten sind mittlerweile verfügbar.[14]) Ein umfassendes und allgemein gültiges Konzept zur Operationalisierung der faktischen Anonymität und zur Überprüfung der Schutzwirkung einer anonymisierten Datei wurde erstmals in Höhne u. a. vorgestellt.[15]) Von besonderer Bedeutung ist dabei auch die Frage, welche Auswirkungen Anonymisierungsverfahren auf den Informationsgehalt und die Ergebnisse von Analysen und somit auf das Analysepotenzial eines Datenbestandes haben. Ein Datensatz, der zwar faktisch anonym ist, gleichzeitig aber nur unbrauchbare

3) Siehe Müller, W./Blien, U./Knoche, P./Wirth, H.: „Die faktische Anonymität von Mikrodaten“, Band 19 der Schriftenreihe „Forum der Bundesstatistik“ des Statistischen Bundesamtes, Wiesbaden 1991.

4) Siehe Brand, R.: „Anonymität von Betriebsdaten – Verfahren zur Erfassung und Maßnahmen zur Verringerung des Reidentifikationsrisikos“, Beiträge zur Arbeitsmarkt- und Berufsforschung, Bd. 237, Nürnberg 2000.

5) Siehe Fußnote 4.

6) Siehe Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik (Hrsg.): „Wege zu einer besseren informationellen Infrastruktur“, Baden-Baden 2001.

7) Siehe Fußnote 6.

8) Siehe Fußnote 6 sowie Vorgrimler, D.: „Aspekte faktischer Anonymisierung“, Arbeitspapier des Projekts „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten“, 2002.

9) Siehe Fußnote 1, hier: S. 472.

10) Siehe Rosemann, M./Vorgrimler, D.: „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten – Strategien, Vorgehen und erste Ergebnisse“, Statistische Analysen, 2004.

11) Siehe Fußnote 2, hier: S. 21.

12) Siehe Fußnote 6.

13) Siehe Lenz, R./Rosemann, M./Vorgrimler, D./Sturm, R.: “Anonymising Business Micro Data – Results of a German Project”, Schmollers Jahrbuch – Journal of Applied Social Science Studies, 126. Jahrgang, Heft 4/2006, S. 635 ff., und Ronning, G./Sturm, R./Höhne, J./Lenz, R./Rosemann, M./Scheffler, M./Vorgrimler, D.: „Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten“, Band 4 der Schriftenreihe „Statistik und Wissenschaft“, Statistisches Bundesamt (Hrsg.), Wiesbaden 2005.

14) Siehe Lenz, R./Vorgrimler, D./Rosemann, M.: „Ein Scientific-Use-File der Kostenstrukturerhebung im Verarbeitenden Gewerbe“ in WiSta 2/2005, S. 91 ff.; Scheffler, M.: „Ein Scientific-Use-File der Einzelhandelsstatistik 1999“ in WiSta 3/2005, S. 197 ff.; Sturm, R./Lenz, R.: „Erste Scientific-Use-Files aus den Wirtschaftsstatistiken“, Proceedings der Nutzerkonferenz am Institut für Weltwirtschaft der Universität Kiel, 19. Mai 2005, S. 191 ff., und Vorgrimler, D./Dittrich, S./Lenz, R./Rosemann, M.: „Ein Scientific-Use-File der Umsatzsteuerstatistik“ in WiSta 3/2005, S. 201 ff.

15) Siehe Höhne, J./Sturm, R./Vorgrimler, D.: „Konzept zur Beurteilung der Schutzwirkung von faktischer Anonymisierung“ in WiSta 4/2003, S. 287 ff.

Ergebnisse liefert, ist für die Wissenschaft wertlos. Dennoch ist klar, dass Informationseinschränkungen auch von Seiten der Nutzer akzeptiert werden müssen, denn jede Form der Anonymisierung ist grundsätzlich mit einer Informationseinschränkung und damit mit einem Verlust an Analysepotenzial verbunden.[16]) Allerdings muss abgewogen werden, welche Formen der Informationseinschränkung aus Sicht der Nutzer mehr oder weniger akzeptabel sind.

Der Beitrag beschäftigt sich mit einigen Aspekten der Beeinträchtigung des Analysepotenzials durch datenverändernde Anonymisierungsverfahren. Betrachtet werden die Auswirkungen von stochastischen Überlagerungen und Mikroaggregationsverfahren auf die Schätzung linearer und nichtlinearer Modelle.

Kapitel 2 gibt zunächst einen Überblick über die wichtigsten datenverändernden Anonymisierungsverfahren. Kapitel 3 beschreibt ausführlicher die im Beitrag betrachteten Verfahren: stochastische Überlagerung und Mikroaggregation. Kapitel 4 beschäftigt sich mit den Auswirkungen von stochastischen Überlagerungen in linearen und nichtlinearen Modellen. Kapitel 5 wendet sich der Verfahrensgruppe der Mikroaggregation und ihren Auswirkungen auf lineare und nichtlineare Modelle zu. Kapitel 6 enthält Praxisbeispiele mit Daten der Kostenstrukturerhebung im Verarbeitenden Gewerbe und des IAB-Betriebspanels. Kapitel 7 fasst abschließend die wichtigsten Ergebnisse zusammen.

2 Überblick über die wichtigsten datenverändernden Anonymisierungsverfahren

Im Laufe der Jahre und Jahrzehnte hat sich eine Vielzahl verschiedener Maßnahmen zur Anonymisierung von Daten entwickelt.[17]) Die Verfahren können grundsätzlich danach unterschieden werden, ob die Anonymisierung in der Einschränkung oder in der Veränderung von Informationen besteht.[18])

Informationsreduzierende Verfahren werden in der Regel bereits bei der Anonymisierung von Haushalts- und Personendaten verwendet. Sie werden daher auch gerne als traditionelle Anonymisierungsverfahren bezeichnet.[19]) Informationsreduktionen werden realisiert, indem Informationen unterdrückt oder vergröbert werden. Die wichtigsten hierbei zu nennenden Verfahren sind:

- Systematische Einschränkung der Grundgesamtheit
- Entfernen auffälliger Merkmalsträger
- (Sub-)Stichprobenziehung
- Variablenunterdrückung
- Vergröberung von Merkmalsausprägungen

Bei informationsverändernden Verfahren werden die Merkmalswerte hingegen systematisch oder stochastisch verändert. Die Verfahren werden deshalb auch als datenverändernde Anonymisierungsverfahren bezeichnet.[20]) Die wichtigsten datenverändernden Verfahren sind:

- Vertauschungsverfahren (Swapping)
- Post-Randomisierung
- Simulationsverfahren
- Imputationsverfahren
- Stochastische Überlagerung
- Mikroaggregation

Swapping-Verfahren weisen den Nachteil auf, dass sie zwar die univariaten Verteilungen exakt erhalten, die Korrelationsbeziehungen zwischen mehreren Variablen jedoch zerstören. Sie sind daher für die Erstellung eines Scientific-Use-File tendenziell eher nicht geeignet.[21])

Beim Verfahren der Post-Randomisierung (PRAM) werden diskrete Merkmale durch die Definition von Übergangswahrscheinlichkeiten randomisiert.[22]) Das Verfahren entspricht der bei Erhebungen verwendeten Randomisierung von Antworten, die durchgeführt wird, um zu erreichen, dass die Befragten auch auf sensible Fragen antworten, beispielsweise nach dem Drogenkonsum oder einer Aids-Erkrankung.[23]) Särndal u. a. haben vorgeschlagen, diese Methode zu verwenden, um die Anonymität von Individuen zu schützen.[24]) Während im Fall randomisierter Antworten das stochastische Modell allerdings definiert werden muss, bevor die Daten erhoben werden, wird bei der Post-Randomisierung die Methode auf die bereits erhobenen Daten angewendet.[25])

Eine Modifikation der Post-Randomisierung – das invariante PRAM – wurde von Höhne (2003) vorgeschlagen und in Ronning und Rosemann (2004) sowie in Ronning u. a. (2005)

16) Siehe Fußnote 6, S. 163.

17) Siehe Fußnote 4 sowie Höhne, J.: „Methoden zur Anonymisierung wirtschaftsstatistischer Einzeldaten“ in Ronning, G./Gnoss, R. (Hrsg.): „Anonymisierung wirtschaftsstatistischer Einzeldaten“, Band 42 der Schriftenreihe „Forum der Bundesstatistik“, Wiesbaden 2003, S. 69 ff.; Ronning, G. u. a., a. a. O., Fußnote 13, Kapitel 4 bis 6, und Rosemann, M.: „Auswirkungen datenverändernder Anonymisierungsverfahren auf die Analyse von Mikrodaten“, IAW-Forschungsbericht Nr. 66, Kapitel 4 bis 6.

18) Siehe Höhne, J., a. a. O., Fußnote 17.

19) Siehe Fußnote 3, Fußnote 4 sowie Ronning, G. u. a., a. a. O., Fußnote 13, Kapitel 4.

20) Siehe Rosemann, M., a. a. O., Fußnote 17, Kapitel 4.

21) Siehe Ronning, G. u. a., a. a. O., Fußnote 13.

22) Siehe Kooiman, P./Willenborg, L./Gouweleeuw, J.: “PRAM: A Method for Disclosure Limitation of Microdata”, Department of Statistical Methods, Statistics Netherlands, Voorburg 1997, und Willenborg, L./de Waal, T.: “Elements of Statistical Disclosure Control”, Lecture Notes in Statistics, Bd. 155, 2001.

23) Siehe Warner, S.: “Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias”, Journal of the American Statistical Association, Bd. 57, 1965, S. 622 ff.

24) Siehe Särndal, C.-E./Swensson, B./Wretman, J.: “Model Assisted Survey Sampling”, New York 1992, S. 573.

25) Siehe Van den Hout, A./van der Heijden, P.: “Randomized Response, Statistical Disclosure Control and Misclassification: A Review”, International Statistical Review, Bd. 70, 2002, S. 269 ff.

formal dargestellt.[26]) Beim invarianten PRAM werden die Randverteilungen erhalten. Der Erwartungswert der anonymisierten Variablen entspricht dem Erwartungswert der Originalvariablen. In Ronning (2005), Ronning und Rosemann (2004) sowie Ronning u. a. (2005) werden die Auswirkungen der Post-Randomisierung der binären abhängigen Variablen auf das Probit-Modell theoretisch abgeleitet.[27]) Dabei wird ein PRAM-korrigierter Probit-Schätzer dargestellt. Ronning (2004) analysiert die Wirkung der Post-Randomisierung einer Dummy-Variablen auf die Varianz- und Kovarianzanalyse.[28]) Biewen[29]) zeigt, wie in einem linearen Modell mit einer post-randomisierten erklärenden nominalen Variablen (Varianzanalyse) unter bestimmten Bedingungen die Schätzung mit Hilfe der verallgemeinerten Momentenmethode korrigiert werden kann.[30])

Simulationsverfahren sind theoretisch optimal, weil im Idealfall synthetische Datensätze sowohl den Vorgaben des Datenschutzes als auch den Anforderungen der Datennutzer genügen.[31]) Die Testdaten lassen sich daher im Idealfall nicht mehr auf die Originaldaten zurückführen.[32]) Auf der anderen Seite werden den Datennutzern für die Untersuchung von kausalen Zusammenhängen und die Schätzung von Populationsmerkmalen brauchbare anonymisierte Daten zur Verfügung gestellt.[33]) Die optimale Umsetzung dieser Idee bestünde darin, die Kerndichte des gesamten Datenbestandes zu schätzen.[34]) Mit Hilfe dieser Dichte würde dann die gewünschte Anzahl der synthetischen Datensätze erzeugt. Allerdings ist die Schätzung mehrdimensionaler empirischer Verteilungen bisher nur für niedrig-dimensionale Daten gelungen. Bereits zwei- bis dreidimensionale Kerndichteschätzungen scheitern daran, dass zu wenige Beobachtungen vorliegen. Folglich können nicht alle Informationen des Originaldatensatzes vollständig im simulierten Datensatz erhalten bleiben.[35]) Dennoch existieren eine Reihe von Simulationsansätzen mit dem Ziel, dem Optimum möglichst nahe zu kommen, beispielsweise das Latin Hypercube Sampling[36]) und das Resampling[37]).

Eine besondere Variante von Simulationsverfahren stellt die Imputation dar. Imputationsverfahren bestehen in einem Austausch von Angaben durch eingeschätzte Werte. Diese Idee wurde zuerst von Rubin[38]) vorgeschlagen und baut auf den Imputationsverfahren im Falle von fehlenden Antworten („Missing Values“) auf, die im Rahmen der Nonresponseforschung entwickelt wurden.[39]) Im Unterschied zur klassischen Anwendung der Imputation bei „Missing Values“ werden hier nicht fehlende Angaben, sondern besonders sensible Merkmalswerte oder Merkmalswerte von Schlüsselvariablen[40]) durch die eingeschätzten Werte ersetzt. Dabei können einzelne Merkmalswerte, die Merkmalswerte besonders gefährdeter Merkmalsträger oder alle Merkmalswerte einzelner bzw. auch aller Variablen anonymisiert werden.

Es wird zwischen einfacher Imputation (single imputation) und multipler Imputation (multiple imputation) unterschieden. Während bei der einfachen Imputation die Einschätzung auf Basis eines einmal unter Einbeziehung aller vorhandenen Beobachtungen geschätzten Regressionsmodells vorgenommen wird, werden bei der multiplen Imputation Bootstrap-Schätzer ermittelt, indem die Regressionsschätzung mit *k* Bootstrap-Stichproben durchgeführt wird.[41])

Sowohl die oben genannten Simulationsverfahren als auch die Imputationsverfahren werden im Folgenden nicht betrachtet, stattdessen erfolgt eine Konzentration auf Mikroaggregationsverfahren und stochastische Überlagerungen. Diese beiden Verfahrensgruppen werden im Folgenden ausführlicher dargestellt.

3 Stochastische Überlagerung und Mikroaggregation

3.1 Stochastische Überlagerung

Die stochastischen Überlagerungen unterteilen sich grundsätzlich in additive und multiplikative Überlagerungen. Variiert werden kann auch die Verteilung des Zufallsfehlers. Additive Zufallsfehler sind in der Regel normalverteilt mit einem Erwartungswert von Null. Neben der Überlagerung mit einer einfachen Normalverteilung ist jedoch auch die Überlagerung

26) Siehe Fußnote 18; Ronning, G./Rosemann, M.: “Estimation of the Probit Model from Anonymised Data”, Beitrag zum Workshop “Econometric Analysis of Anonymised Firm Data”, Tübingen, März 2004, und Ronning, G./Rosemann, M./Strotmann, H.: “Post-Randomization under Test: Estimation of the Probit Model”, Jahrbücher für Nationalökonomie und Statistik, Bd. 225(5), 2005, S. 544 ff.

27) Siehe Ronning, G.: “Randomized Response and the Binary Probit Model”, Economics Letters, Bd. 86, 2005, S. 221 ff.; Ronning, G./Rosemann, M., a. a. O., Fußnote 26, und Ronning, G./Rosemann, M./Strotmann, H., a. a. O., Fußnote 26.

28) Siehe Ronning, G.: „Fehlklassifikation im Modell der Varianzanalyse“, Arbeitspapier des Projekts „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten“, 2004.

29) Siehe Biewen, E.: „Mikroökonometrische Evaluation bei Fehlklassifikation der Treatment-Variablen“, unveröffentlichte Diplomarbeit, Universität Tübingen, 2005.

30) Ein Überblick über die Wirkung der Post-Randomisierung in deskriptiven Auswertungen und im Probit-Modell findet sich in Ronning, G. u. a., a. a. O., Fußnote 13. Auch wird eine Kombination aus Randomisierung der abhängigen binären Variablen und stochastischer Überlagerung bzw. Mikroaggregation der erklärenden metrischen Variablen im Probit- Modell behandelt.

31) Siehe Gottschalk, S.: „Unternehmensdaten zwischen Datenschutz und Analysepotenzial“, Baden-Baden 2005, S. 115.

32) Statistische Ämter des Bundes und der Länder und IAW: „Forschungsprojekt: ‚Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten‘ – Zwischenbericht 2003 an das BMBF“, Statistisches Bundesamt, Wiesbaden.

33) Siehe Fußnote 31.

34) Siehe Fienberg, S.: “Confidentiality and Disclosure Limitation Methodology: Challenges for National Statistics and Statistical Research”, Technical Report No. 668, Pittsburgh 1997.

35) Siehe Fußnote 31.

36) Siehe Dandekar, R./Cohen, M./Kirkendall, N.: “Applicability of Latin Hypercube Sampling to Create Multivariate Synthetic Micro Data” in Proceedings of ETK-NTTS, Eurostat, Luxemburg 2001, S. 839 ff.

37) Siehe Gottschalk, S.: “Microdata Disclosure Control by Resampling – Empirical Findings for Business Survey Data” in Allgemeines Statistisches Archiv, Bd. 88(3), 2004, S. 279 ff., und Gottschalk, S., a. a. O., Fußnote 31.

38) Siehe Rubin, D.: “Discussion: Statistical Disclosure Limitation”, Journal of Official Statistics, Bd. 9(2), 1993, S. 461 ff.

39) Siehe Fußnote 34.

40) Überschneidungsmerkmale mit dem Zusatzwissen.

41) Siehe Little, R.: “Statistical Analysis of Masked Data”, Journal of Official Statistics, Bd. 9, 2003, S. 407 ff.; Raghunathan, T./Reiter, J./Rubin, D.: “Multiple Imputation for Statistical Disclosure Limitation”, Journal of Official Statistics, Bd. 19, 2003, S. 1 ff., und Rubin, D./Schenker, N.: “Multiple Imputation in Health-Care Databases: An Overview and some Applications”, Statistics in Medicine, Bd. 10, 1991, S. 585 ff.

mit einem Zufallsfehler möglich, der aus einer Mischungsverteilung aus mehreren Normalverteilungen gezogen wird. Die gleichen Verteilungen sind auch bei multiplikativen Überlagerungen verwendbar. Allerdings muss dann eine Verteilung mit ausschließlich positiven Werten und Erwartungswert Eins gewählt werden. Alternativ kann der multiplikative Zufallsfehler auch gleichverteilt im positiven Bereich sein, einer Lognormalverteilung oder einer gestutzten Normalverteilung entstammen.

a) Additive stochastische Überlagerung

Bei einer additiven stochastischen Überlagerung mit einer Normalverteilung werden die einzelnen Merkmalswerte mit einem Zufallsfehler überlagert, dessen Erwartungswert den Wert Null aufweist und dessen Varianz beziehungsweise Varianz-Kovarianzmatrix konstant ist. Wird die Varianz-Kovarianzmatrix der Überlagerungen proportional zur Varianz-Kovarianzmatrix der Originalwerte gewählt, so ist auch die Varianz-Kovarianzmatrix der anonymisierten Daten proportional zur Varianz-Kovarianzmatrix der Originalwerte. Damit entspricht die Korrelationsmatrix der anonymisierten Werte derjenigen der Originalwerte (im theoretischen Sinne). In linearen Regressionsmodellen werden die Regressionsschätzer somit asymptotisch erwartungstreu, sofern alle Variablen (einschließlich der abhängigen) additiv überlagert werden. Die Varianz der Störgrößen wird jedoch um den Faktor $1 + d$ überschätzt.[42]) Kim schlägt vor, dies durch eine zusätzliche Transformation zu korrigieren (Kim-Verfahren).[43])

Ein Problem der additiven stochastischen Überlagerung mit einer einfachen Normalverteilung mit Erwartungswert Null besteht darin, dass der Zufallsfehler mit einer hohen Wahrscheinlichkeit Werte nahe Null annimmt, die überlagerten Werte folglich auch mit einer hohen Wahrscheinlichkeit nahe bei den Originalwerten liegen. Will man das ändern, so kann man eine höhere Varianz verwenden. Dies birgt allerdings das Risiko, dass einzelne Werte sehr stark von den entsprechenden Originalwerten abweichen. Deshalb hat Roque vorgeschlagen, bei der Überlagerung anstatt einer einfachen Normalverteilung eine Mischung aus normalverteilten Zufallswerten zu nutzen.[44]) Ähnliche Vorschläge finden sich bei Höhne und Yancey u. a.[45]) Damit kann bei gleicher Varianz erreicht werden, dass ein größerer Anteil der überlagerten Werte weiter von den Originalwerten entfernt ist. Dieser Effekt kann bereits bei zwei Mischungskomponenten erreicht werden.[46])

b) Multiplikative stochastische Überlagerung

Die multiplikative stochastische Überlagerung weist gegenüber der additiven den Vorteil auf, dass der stärkeren Reidentifikationsgefährdung größerer Unternehmen Rechnung getragen wird. Zudem erhält sie die Nullen und, sofern ausschließlich positive Überlagerungsfaktoren verwendet werden, auch die Vorzeichen.

Bei der multiplikativen Überlagerung besteht die Möglichkeit, die Daten eines Merkmalsträgers entweder mit einem konstanten Zufallsfaktor zu überlagern oder für jedes Merkmal einen neuen Zufallsfaktor zu erzeugen.[47]) Die erste Vorgehensweise hat aus Nutzersicht den Vorteil, dass die relativen Beziehungen zwischen den Variablen eines Merkmalsträgers erhalten bleiben. Dies kann allerdings auch zu einem höheren Reidentifikationsrisiko führen.

Wie bereits erwähnt, sollten die Überlagerungsfaktoren lediglich positive Werte annehmen, damit die Vorzeichen der Merkmalsträger nicht systematisch verändert werden. Aus diesem Grund werden bei multiplikativen Überlagerungen für die Störgrößen in der Regel Verteilungen verwendet, die lediglich positive Ausprägungen der Merkmalswerte zulassen, beispielsweise eine Gleichverteilung im positiven Wertebereich[48]) oder die Lognormalverteilung[49]).

Um lediglich positive Werte für die Zufallsfehler zu erhalten, kann auch, wie ebenfalls von Kim und Winkler vorgeschlagen, eine gestutzte Normalverteilung verwendet werden. Höhne schlägt hingegen vor, die Varianzen der Zufallsfehler so gering zu wählen, dass Werte kleiner als oder gleich Null für die Überlagerungsfaktoren bei einem Erwartungswert von Eins auch bei einer Normalverteilung nur mit sehr kleiner Wahrscheinlichkeit auftreten. Treten sie in der Praxis dennoch auf, werden die Zufallsfehler erneut erzeugt. Um dennoch einen ausreichenden Schutz der Daten zu gewährleisten, soll analog zum additiven Fall eine zweigipflige Mischungsverteilung für den Zufallsfehler verwendet werden.

Eine spezielle Form einer Mischungsverteilung wird von Höhne zur Anonymisierung vorgeschlagen: Bei diesem, im Folgenden auch als Höhne-Verfahren bezeichneten Vorgehen wird zunächst mit einer Wahrscheinlichkeit von 0,5 festgelegt, ob die Merkmalswerte eines Merkmalsträgers verkleinert oder vergrößert werden. Hierzu werden die Grundüberlagerungsfaktoren $1 - f$ und $1 + f$ festgelegt. Jeder Merkmalsträger erhält einen dieser Grundüberlagerungsfaktoren (mit Wahrscheinlichkeit 0,5) zugewiesen. Die Grundüberlagerungsfaktoren werden anschließend für jeden Merkmalswert unabhängig additiv mit einer Normalverteilung mit Erwartungswert Null und Standardabweichung s ($s < f/2$) überlagert. Somit wird jeder Merkmalsträger in die gleiche Richtung verzerrt, dennoch wird jeder Merkmalswert

42) Siehe Fußnote 4.

43) Siehe Kim, J.: "A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation", Proceedings of the Section on Survey Research Methods, American Statistical Association, 1986, S. 370 ff.

44) Siehe Roque, G.: "Masking Microdata Files with Mixtures of Multivariate Normal Distributions", Dissertation, Riverside 2004.

45) Siehe Höhne, J.: „Varianten von Zufallsüberlegungen", Arbeitspapier des Projekts „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten", 2004, und Yancey, W./Winkler, W./Creezy, R.: "Disclosure Risk Assessment in Perturbative Micro Data Protection" in Domingo-Ferrer, J. (Hrsg.): "Inference Control in Statistical Database – From Theory to Practice", Berlin 2002, S. 135 ff.

46) Zu Details siehe auch Ronning, G. u. a., a. a. O., Fußnote 13, und Rosemann, M., a. a. O., Fußnote 17.

47) Siehe Höhne, J., a. a. O., Fußnote 45.

48) Siehe Fußnote 31.

49) Siehe Kim, J./Winkler, W.: "Multiplicative Noise for masking Continuous Data", Proceedings of the Section on Survey Research Methods, American Statistical Association, 2001.

eines Merkmalsträgers mit einem anderen Überlagerungsfaktor multipliziert.

Insbesondere bei sehr schief verteilten Originalvariablen hängt die Stärke der Abweichungen durch Überlagerung von der Konstellation der Zufallszahlen bei wenigen großen Merkmalsträgern ab. Dadurch kann es passieren, dass trotz der asymptotischen Erwartungstreue und qualitativ hochwertig generierten Zufallszahlen die Mittelwerte und Summen nur sehr schlecht reproduziert werden. Höhne entwickelt deshalb einen Algorithmus für eine „kontrollierte" multiplikative Überlagerung. Dabei wird sichergestellt, dass auch die Mittelwerte innerhalb bestimmter Teilbereiche annähernd erhalten bleiben.

3.2 Mikroaggregationsverfahren

Die Grundidee der Mikroaggregationsverfahren besteht darin, ähnliche Objekte zu Gruppen zusammenzufassen und die Ursprungswerte durch die arithmetischen Mittel der Merkmalswerte aller Merkmalsträger innerhalb der Gruppen zu ersetzen.[50]) Alle Gruppierungsverfahren gehen von Gruppen mit mindestens drei Werten aus, denn bei nur zwei Merkmalsträgern können die Werte des einen Merkmalsträgers bei Kenntnis der Werte des anderen Merkmalsträgers in jedem Fall enthüllt werden.

Grundsätzlich kann zwischen zwei Arten der Mikroaggregation unterschieden werden:

- Deterministische bzw. abstandsorientierte Mikroaggregation, bei der möglichst ähnliche Einheiten zusammengefasst werden.
- Stochastische Mikroaggregation, bei der die Gruppenbildung zufällig erfolgt.

Zudem erfolgt eine Unterscheidung danach, ob die Mikroaggregation für alle Variablen gemeinsam erfolgt – für die Durchschnittsbildung bei den verschiedenen Variablen folglich die gleichen Gruppen gebildet werden – oder die Gruppenbildung für jede Variable getrennt erfolgt. Abstandsorientierte Mikroaggregationsverfahren unterscheiden sich zusätzlich nach dem verwendeten Abstandsmaß.[51])

a) Abstandsorientierte Mikroaggregation

Die Idee der deterministischen beziehungsweise abstandsorientierten Mikroaggregation besteht darin, möglichst ähnliche Merkmalsträger zu Gruppen zusammenzufassen und deren Originalwerte durch die arithmetischen Mittel innerhalb der Gruppen zu ersetzen. Die einzelnen Verfahrensvarianten unterscheiden sich zum einen danach, ob die Gruppenbildung für alle metrischen Variablen – oder auch Gruppen von Variablen – gemeinsam erfolgt (mehrdimensionale Mikroaggregation) oder die Variablen getrennt mikroaggregiert werden (eindimensionale Mikroaggregation). Zum anderen unterscheiden sich die mehrdimensionalen Mikroaggregationsverfahren hinsichtlich der Bestimmung des Abstandes zwischen den einzelnen Objekten.

a1) Abstandsorientierte Mikroaggregation für alle Variablen gemeinsam (gemeinsame Mikroaggregation)

- **Mikroaggregation nach einer Variablen:** Es wird eine dominierende Variable herausgesucht und der Datenbestand danach sortiert. Danach werden absteigend immer drei benachbarte Merkmalsträger in einer Gruppe zusammengefasst und alle ihre stetigen Merkmalswerte durch den Durchschnitt der Werte der Gruppe ersetzt. (Die dominierende Variable sollte dabei mit möglichst vielen weiteren Merkmalen stark korreliert sein.)
- **Mikroaggregation nach einer Hilfsvariablen:** Die Sortierung erfolgt anhand von Hilfsvariablen. Die Hilfsvariablen sind dabei zum Beispiel die Hauptkomponente (als eine durch Transformation gebildete Variable mit möglichst hoher Korrelation zu den anderen Variablen) oder die Z-Scores (als die Summe der standardisierten Originalvariablen).
- **Mikroaggregation nach allen metrischen Variablen:** Die Gruppenbildung erfolgt beispielsweise nach der euklidischen Distanz zwischen den Merkmalsträgern. Dabei werden die beiden Merkmalsträger herausgesucht, die den größten Abstand untereinander haben. Danach werden diesen beiden jeweils die zwei dichtesten Merkmalsträger hinzu gruppiert. Die verbleibenden, noch nicht gruppierten Merkmalsträger werden wieder analog behandelt.[52])

a2) Abstandsorientierte Mikroaggregation für alle Variablen getrennt (getrennte Mikroaggregation)

Der Datenbestand wird jeweils nach der zu anonymisierenden Variablen sortiert. Danach werden absteigend immer drei bis fünf benachbarte Merkmalsträger in einer Gruppe zusammengefasst und alle ihre stetigen Merkmalswerte für die betrachtete Variable durch den Durchschnitt der Werte der Gruppe ersetzt. Anschließend wird der Vorgang für die anderen metrischen Variablen wiederholt. Die Sortierung erfolgt demnach für jede Variable neu.

a3) Abstandsorientierte Mikroaggregation für Gruppen von Variablen (gruppierte Mikroaggregation/teilweise gemeinsame Mikroaggregation)

Bei dieser Verfahrensvariante werden die Variablen zunächst gruppiert und anschließend innerhalb der gebildeten Gruppen gemeinsam mikroaggregiert. Diese Variante der Mikroaggregation wurde von Domingo-Ferrer und Mateo-Sanz entwickelt.[53]) Die Gruppenbildung der Variablen erfolgt nach den Korrelationen zwischen den Variablen.[54]) Für die einzelnen Variablengruppen kann die Mikroaggregation ana-

50) Siehe Mateo-Sanz, J. M./Domingo-Ferrer, J.: "A Method for Data-Oriented Multivariate Microaggregation" in Statistical Data Protection, Proceedings of the Conference, Eurostat 1999.
51) Siehe Schmid, M.: "Estimation of a Linear Regression with Microaggregated Data", München 2007.
52) Siehe Fußnote 50.
53) Siehe Domingo-Ferrer, J./Mateo-Sanz, J. M.: "An Empirical Comparison of SDC Methods for Continuous Microdata in Terms of Information Loss and Disclosure Risk", Second Eurostat-UN/ECE Joint Work Session on Statistical Data Confidentiality, Skopje 2001.
54) Siehe Fußnote 32.

log der in a1) beschriebenen Varianten für die gemeinsame Mikroaggregation vorgenommen werden. Die abstandsorientierte Gruppenbildung der Objekte kann somit nach einer Variablen, nach allen metrischen Variablen innerhalb der Gruppe oder nach aus diesen gebildeten Hilfsvariablen erfolgen.

b) Stochastische Mikroaggregation

Stochastische Mikroaggregationsverfahren wurden erstmals von Lechner und Pohlmeier vorgeschlagen.[55]) Sie beschreiben zwei Möglichkeiten der stochastischen Mikroaggregation, die im Folgenden als zufällige Mikroaggregation und als Bootstrap-Mikroaggregation bezeichnet werden.

- Das Vorgehen bei der zufälligen Mikroaggregation entspricht grundsätzlich dem Vorgehen bei der deterministischen Mikroaggregation, allerdings erfolgt die Gruppenbildung der Objekte nicht abstandsorientiert, sondern zufällig. Damit spielt die Ähnlichkeit der Objekte bei der Gruppenbildung keine Rolle. Die zufällige Gruppenbildung kann analog zur deterministischen Mikroaggregation für alle Variablen – beziehungsweise für Gruppen von Variablen – gemeinsam oder für alle Variablen getrennt erfolgen.

- Bei der Bootstrap-Mikroaggregation werden für jedes Objekt zufällig zwei weitere gezogen. Die Ziehung erfolgt mit Zurücklegen – auch das erste Unternehmen selbst kann nochmals gezogen werden. Diese drei Objekte bilden eine Gruppe, deren durchschnittliche Merkmalswerte an die Stelle der Werte für das erste Objekt treten.

4 Stochastische Überlagerung in linearen und nichtlinearen Modellen: Ein Überblick

4.1 Stochastische Überlagerung im linearen Modell

Die Auswirkungen der additiven stochastischen Überlagerung im linearen Modell sind aus der Standardliteratur über die Messfehlerproblematik bekannt.[56]) Sofern lediglich die abhängige Variable überlagert wird, bleibt der Kleinst-Quadrate-Schätzer (OLS) erwartungstreu. Es ergibt sich lediglich eine im Vergleich mit den unverzerrten Daten erhöhte Residuenvarianz. Werden hingegen die Regressoren – oder nur ein Teil von ihnen – mit einem additiven Fehler überlagert, so ergibt sich eine Verzerrung des OLS-Schätzers. Für den Wahrscheinlichkeitsgrenzwert des Vektors der Koeffizientenschätzer $\hat{\beta}$ ergibt sich in diesem Fall:[57])

$$(1) \qquad p\lim\hat{\beta} = (\mathbf{Q} + \Sigma_{ww})^{-1}\mathbf{Q}\beta \,.$$

Dabei ist $\mathbf{Q} = p\lim\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)$ mit der Regressormatrix $\mathbf{X}$ und Σ_{ww} die Varianz-Kovarianzmatrix der Überlagerungen. Umgeformt ergibt sich das Standardresultat des Fehler-in-den-Variablen-Modells:

$$(2) \qquad p\lim\hat{\beta} = \beta - (\mathbf{Q} + \Sigma_{ww})^{-1}\Sigma_{ww}\beta \,.$$

Werden Regressoren und abhängige Variable additiv überlagert, so hängt der Wahrscheinlichkeitsgrenzwert des Schätzers auch von der Kovarianz-Matrix Σ_{wv} zwischen dem Überlagerungsvektor der abhängigen Variablen v und der Überlagerungsmatrix der Regressoren $\boldsymbol{W}$ ab:

$$(3) \qquad p\lim\hat{\beta} = (\mathbf{Q} + \Sigma_{ww})^{-1}(\mathbf{Q}\beta + \Sigma_{wv}) \,.$$

Gewöhnlich geht man in Fehler-in-den-Variablen-Modellen davon aus, dass die Messfehler unkorreliert sind und der Ausdruck Σ_{wv} vernachlässigbar ist. Wird jedoch die additive stochastische Überlagerung als Anonymisierungsverfahren eingesetzt, so kann diese Annahme bewusst verletzt werden, insbesondere dann, wenn die Varianz-Kovarianzmatrix der Überlagerungen bewusst als ein Vielfaches der Varianz-Kovarianzmatrix der Ausgangsvariablen gewählt wird: $\Sigma_{ww} = d\Sigma_{xx}$ und $\Sigma_{wv} = d\Sigma_{xy}$. Für diesen Fall kann man zeigen, dass der OLS-Schätzer konsistent ist, sofern alle Variablen überlagert werden.[58]) Eine konsistente Schätzung erhält man auch bei Anwendung des Kim-Verfahrens. Das Gleiche gilt, falls nicht alle Regressoren überlagert werden, aber die originale Varianz-Kovarianz-Matrix durch entsprechende Transformationen erhalten bleibt.[59])

Für den Fall inkonsistenter Schätzer lassen sich aus den Formeln (2) und (3) Korrekturschätzer ableiten.[60]) Alternativ hierzu kann der Instrumentvariablen-Schätzer (IV-Schätzer) herangezogen werden. Dabei müssen die Instrumentvariablen asymptotisch unkorreliert mit den Messfehlern und den Überlagerungen sowie hoch korreliert mit den Regressoren sein. Bei der Anonymisierung wirtschaftsstatistischer Mikrodaten kann den Nutzern ein Datensatz mit den Instrumentvariablen dadurch zur Verfügung gestellt werden, dass der Originaldatenbestand zweimal unabhängig überlagert wird und die Datennutzer somit zwei anonymisierte Datensätze erhalten.

Weniger bekannt sind die Auswirkungen von multiplikativen stochastischen Überlagerungen auf die Kleinst-Quadrate-Schätzung im linearen Modell. Auch hier gilt, dass der OLS-Schätzer bei alleiniger Überlagerung der abhängigen Variablen weiterhin konsistent ist. Anders verhält es sich wiederum für den Fall, dass die Regressoren überlagert wer-

55) Siehe Lechner, S./Pohlmeier, W.: „Schätzung ökonometrischer Modelle auf der Grundlage anonymisierter Daten“ in Ronning, G./Gnoss, R. (Hrsg.), a. a. O., Fußnote 17, S. 115 ff.
56) Siehe Fuller, W.: “Properties of Some Estimators for the Errors-in-Variables Model”, The Annals of Statistics, Bd. 8, 1980, S. 407 ff.; Fuller, W.: “Measurement Error Models with Heterogeneous Error Variances” in Chaubey, Y./Dwivedi, T. (Hrsg.): “Topics in Applied Statistics”, 1984, S. 257 ff., und Fuller, W.: “Measurement Error Models”, New York 1987.
57) Zur Herleitung siehe auch Rosemann, M., a. a. O., Fußnote 17.
58) Siehe Fußnote 4.
59) Siehe Muralidhar, K./Parsa, R./Sarathy, R.: “A General Additive Data Perturbation Method for Database Security”, Management Science, Vol. 45, No. 10, 1999, S. 1399 ff.
60) Siehe Fuller, W.: “Measurement Error Models”, New York 1987, sowie Fußnote 55.

den. Werden ausschließlich die Regressoren multiplikativ überlagert, so erhält man für den Wahrscheinlichkeitsgrenzwert des Schätzers:[61])

$$(4) \qquad p\lim\hat{\beta} = (\mathbf{M} \circ \mathbf{Q})^{-1}\mathbf{Q}\beta \,.$$

Dabei ist $\mathbf{M} \circ \mathbf{Q}$ das Hadamard-Produkt der Matrizen **M** und **Q**. Dies bezeichnet die elementweise Multiplikation beider Matrizen. Außerdem gilt $\mathbf{M} = E(\mathbf{w}_i\mathbf{w}_i')$ mit $\mathbf{w}_i'$ der i-ten Zeile der Überlagerungsmatrix **W**.

Werden abhängige Variable und Regressoren multiplikativ überlagert, so ergibt sich für den Wahrscheinlichkeitsgrenzwert des OLS-Schätzers:[62])

$$(5) \qquad p\lim\hat{\beta} = (\mathbf{M} \circ \mathbf{Q})^{-1}(\mathbf{Q}\beta \circ \mathbf{k}) \;\text{ mit }\; \mathbf{k} = E(\mathbf{w}_i v_i) \,.$$

Damit hängt in diesem Fall die Verzerrung des Schätzers wiederum von der Kovarianz zwischen dem Überlagerungsfaktor der abhängigen Variablen *V* und den Überlagerungsfaktoren der Regressoren ab. Sind diese nicht korreliert, so entspricht das Ergebnis aus Gleichung (5) demjenigen aus Gleichung (4).

Falls alle Merkmalswerte einer Einheit mit einem konstanten Faktor überlagert werden, so ergibt sich in zwei speziellen Fällen allerdings ein konsistenter OLS-Schätzer. Zum einen wenn das Absolutglied Null ist, zum anderen wenn für alle Regressoren $\sum_{i=1}^{n} x_i = 0$ gilt.[63])

Für den Fall eines inkonsistenten Schätzers kann eine Korrektur analog zum Fall der additiven stochastischen Überlagerung mit Hilfe des Instrumentvariablen-Schätzers erfolgen, sofern geeignete Instrumentvariablen zur Verfügung stehen. Hwang schlägt für den Fall, dass ausschließlich die Regressoren multiplikativ überlagert werden, folgenden konsistenten Korrekturschätzer vor, der sich aus Gleichung (4) ergibt:[64])

$$(6) \qquad \hat{\beta}^{Hwa} = \left[\left(\mathbf{X}^{a\prime}\mathbf{X}^{a}\right) \div \hat{\mathbf{M}}\right]^{-1}\mathbf{X}^{a\prime}\mathbf{y} \;\text{ mit }\; \hat{\mathbf{M}} = \frac{\mathbf{W}'\mathbf{W}}{n} \,.$$

Dabei bezeichnet ÷ die Hadamard-Division, also die elementweise Division, $\mathbf{X}^a$ die Matrix der anonymisierten (überlagerten) Regressoren.

Der Schätzer kann auch verwendet werden, wenn auch die abhängige Variable multiplikativ überlagert wird, allerdings nur, sofern der Überlagerungsfaktor der abhängigen Variablen mit den Überlagerungsfaktoren der Regressoren unkorreliert ist.[65]) Ist die Bedingung hingegen nicht erfüllt, so ergibt sich aus Gleichung (5) ein Korrekturschätzer durch:[66])

$$(7) \qquad \hat{\beta}^{Hwa2} = \left[\left(\mathbf{X}^{a\prime}\mathbf{X}^{a}\right) \div \hat{\mathbf{M}}\right]^{-1}(\mathbf{X}^{a\prime}\mathbf{y}^{a} \div \hat{\mathbf{k}}) \;\text{ mit }\; \hat{\mathbf{k}} = \frac{\mathbf{W}'\mathbf{v}}{n} \,.$$

Dabei bezeichnet $\mathbf{y}^a$ den Vektor der anonymisierten (überlagerten) abhängigen Variablen. Voraussetzung für die Anwendung der beiden Korrekturschätzer ist neben der zeilenweisen Unkorreliertheit der Überlagerungen auch, dass $\hat{\mathbf{M}} = \frac{\mathbf{W}'\mathbf{W}}{n}$ und gegebenenfalls $\hat{\mathbf{k}} = \frac{\mathbf{W}'\mathbf{v}}{n}$ durch die datenerstellende Institution zur Verfügung gestellt werden oder beide Größen geschätzt werden können.

4.2 Stochastische Überlagerung in nichtlinearen Modellen

Im vorangegangenen Abschnitt wurde hergeleitet, dass stochastische Überlagerungen der Regressoren in der Regel zu inkonsistenten Schätzern im linearen Modell führen. Dies gilt ebenso für die Schätzung von nichtlinearen Modellen. Einen Überblick über die Messfehlerproblematik in nichtlinearen Modellen geben Carroll u. a.[67]) Allerdings sind die Beschaffenheit der Verzerrung und damit auch deren Korrektur in nichtlinearen Modellen weitaus komplexer als in linearen Modellen.

Aufgrund der Verschiedenheit der nichtlinearen Modelle wurden spezielle Korrekturmodelle entwickelt, beispielsweise von Stefanski und Carroll[68]) für die logistische Regression. Daneben existieren jedoch auch generelle Fehler-Korrektur-Schätzer für nichtlineare Modelle[69]), von denen insbesondere der Kalibrationsschätzer und der SIMEX-Schätzer[70]) zu nennen sind, die auch im Programmpaket STATA implementiert sind. SIMEX-Schätzer und Kalibration sind für alle Schätzmethoden anwendbar.[71]) Somit ist – ebenso wie für stochastische Überlagerungen im linearen Modell – auch in nichtlinearen Modellen eine Korrektur möglich. Im Folgenden wird die SIMEX-Methode ausführlicher vorgestellt, da dieser ein sehr anschauliches Vorgehen zugrunde liegt.

Die SIMEX-Methode wurde von Cook und Stefanski[72]) zur Korrektur von Messfehlern in nichtlinearen Modellen vorgeschlagen. SIMEX steht für „Simulation Extrapolation".

61) Siehe Hwang, J.: "Multiplicative Errors-in-Variables Models with Applications to Recent Data Released by the U.S. Department of Energy", Journal of the American Statistical Association, Bd. 81(395), 1986, S. 680 ff., und Lin, A.: "Estimation of Multiplicative Measurement Error Models and Some Simulation Results", Economics Letters, Bd. 31, 1989, S. 13 ff.

62) Beide Ergebnisse setzen voraus, dass die Überlagerungen unterschiedlicher Einheiten unkorreliert sind (zeilenweise unkorreliert). Die Herleitungen dieser Ergebnisse finden sich im Anhang von Rosemann, M., a. a. O., Fußnote 17.

63) Siehe Rosemann, M., a. a. O., Fußnote 17.

64) Siehe Hwang, J., a. a. O., Fußnote 61.

65) Siehe Lin, A., a. a. O., Fußnote 61.

66) Siehe Rosemann, M., a. a. O., Fußnote 17.

67) Siehe Carroll, R./Ruppert, D./Stefanski, L.: "Measurement Error in Nonlinear Models", London 1995.

68) Siehe Stefanski, L./Carroll, R.: "Covariate Measurement Error in Logistic Regression", The Annals of Statistics, Bd. 13, 1985, S. 1335 ff.

69) Siehe Fußnote 67.

70) Siehe hierzu auch Lechner, S./Pohlmeier, W.: "Data Masking by Noise Addition and the Estimation of Nonlinear Regression Models", Jahrbücher für Nationalökonomie und Statistik, Bd. 225(5), 2005, S. 517 ff.

71) Siehe Fußnote 67.

72) Siehe Cook, J./Stefanski, L.: "Simulation-Extrapolation Estimation in Parametric Measurement Error Models", Journal of the American Statistical Association, Bd. 89(428), 1994, S. 1314 ff.

Die Methode kann angewendet werden, sofern die Varianz der Überlagerungen bekannt ist oder gut geschätzt werden kann, bei anonymisierten Daten beispielsweise, indem ein zweiter in gleicher Weise (gleiche Verteilungsfamilie, gleicher Mittelwert, gleiche Varianz der Überlagerungen) stochastisch überlagerter Datensatz zur Verfügung gestellt wird.

Der SIMEX-Schätzer wurde für den Fall eines additiven Messfehlers beziehungsweise für die additive stochastische Überlagerung entwickelt. Die Idee des SIMEX-Schätzers lässt sich zunächst am besten anhand des linearen Regressionsmodells mit einem Regressor *X* beschreiben. Der Wahrscheinlichkeitsgrenzwert des OLS-Schätzers ohne Korrektur der Überlagerung lautet in diesem Fall:

$$(8) \qquad p\lim \hat{\beta}_x = \frac{\beta_x \sigma_x^2}{\sigma_x^2 + \sigma_v^2}.$$

Dabei ist σ_x^2 die Varianz des Regressors, σ_v^2 die Varianz der Überlagerung.

Dem SIMEX-Schätzer liegt nun die Idee zugrunde, die Effekte der stochastischen Überlagerung experimentell mit Hilfe von Simulationen zu bestimmen. Dabei wird der Zusammenhang zwischen der Stärke der Überlagerung und dem naiven Schätzer bestimmt. Anschließend wird mit Hilfe des gefundenen Zusammenhangs auf den Zustand ohne Überlagerung zurückgeschlossen.

Ist die Varianz der additiven Überlagerung bekannt, so kann die anonymisierte Variable mit einem zusätzlichen Fehler überlagert werden, dessen Varianz das λ-Fache der ursprünglichen Fehlervarianz beträgt. Damit ergibt sich für den Wahrscheinlichkeitsgrenzwert des naiven Schätzers in Abhängigkeit von λ:

$$(9) \qquad p\lim \hat{\beta}_x = \frac{\beta_x \sigma_x^2}{\sigma_x^2 + (1+\lambda)\sigma_v^2}.$$

Variiert man nun den Parameter λ, so erhält man den gewünschten funktionalen Zusammenhang zwischen der Stärke der Überlagerung und dem naiven Schätzer. Der wahre Wert für den Schätzer würde sich nun ergeben, wenn λ den Wert -1 annehmen würde. Deshalb kann der wahre Schätzwert durch eine Extrapolation des funktionalen Zusammenhangs zwischen λ und $\hat{\beta}_x$ auf $\lambda = -1$ ermittelt werden. Diese Überlegung wird bei der Berechnung des SIMEX-Schätzers auf nichtlineare Zusammenhänge übertragen.

Das Vorgehen bei der Berechnung des SIMEX-Schätzers besteht aus zwei Schritten, dem Simulationsschritt und dem Extrapolationsschritt. Im Simulationsschritt wird für $\lambda > 0$ eine zusätzliche Überlagerung der anonymisierten Merkmalswerte vorgenommen. Dies erfolgt für jede Stufe λ in *B* Wiederholungen. Somit können auch für jedes λ insgesamt *B* naive Schätzer berechnet werden. Das arithmetische Mittel hieraus ergibt den Schätzer in Abhängigkeit von λ.

Der zweite und schwierigere Schritt besteht in der Extrapolation. Als besondere Herausforderung erweist sich dabei die Wahl einer geeigneten Extrapolationsfunktion. Die Extrapolationsfunktion muss so gewählt werden, dass sie die im ersten Schritt berechneten Schätzer $\hat{\beta}_x(\lambda)$ optimal verbindet und den wahren Schätzer an der Stelle $\lambda = -1$ möglichst genau trifft. Cook und Stefanski halten drei Formen der Extrapolationsfunktion für denkbar:

— eine lineare Extrapolationsfunktion:

$$G_L(\lambda) = \gamma_1 + \gamma_2 \lambda.$$

— eine quadratische Extrapolationsfunktion:

$$G_Q(\lambda) = \gamma_1 + \gamma_2 \lambda + \gamma_3 \lambda^2.$$

— eine nichtlineare Extrapolationsfunktion:

$$G_{RL}(\lambda) = \gamma_1 + \frac{\gamma_2}{\gamma_3 + \lambda}.$$

Sofern die Überlagerung normalverteilt ist, ist jede dieser drei Extrapolationsfunktionen exakt für verschiedene Schätzer. Die Auswahl der geeigneten Extrapolationsfunktion ist wohl das größte Problem bei der Anwendung des SIMEX-Schätzers. Sie ist deshalb sorgfältig zu treffen. Zu beachten ist, dass bei der rational linearen Funktion numerische Instabilitäten auftreten können. Auf der anderen Seite sind sowohl die lineare als auch die quadratische Extrapolationsfunktion hinsichtlich der Korrektur der Schätzung konservativer. Der SIMEX-Schätzer ist generell lediglich approximativ konsistent, weil es sich bei der geschätzten Extrapolationsfunktion ebenfalls um eine Approximation handelt. Lediglich in den Spezialfällen, in denen die Extrapolationsfunktion exakt ist, ist auch der SIMEX-Schätzer exakt konsistent.

Generell gilt, dass der SIMEX-Schätzer nicht auf den Fall normalverteilter Überlagerungen beschränkt ist. Auch ist er nicht auf den Fall der additiven Überlagerung beschränkt. Die überlagerte Variable kann durch eine Transformation *H* in ein Modell mit additiver Überlagerung überführt werden, sodass $H(X^a) = H(X) + V$ gilt. Notwendig ist, dass zu *H* eine inverse Funktion *F* existiert. Im Fall der multiplikativen Überlagerung gilt $H = \log(\,)$ und $F = \exp(\,)$.

Allerdings kann die multiplikative Überlagerung mit einem Überlagerungsfaktor *W* auch behandelt werden, als ob es sich um eine additive Überlagerung mit *V* handeln würde. Dies kann wie folgt veranschaulicht werden: Im Fall der additiven Überlagerung (mit einem Erwartungswert von Null) ergibt sich Erwartungstreue sowie für die Varianz der anonymisierten Variablen:

$$(10) \qquad \mathrm{var}\left(X^{a(add)}\right) = \sigma_x^2 + \sigma_v^2.$$

Für eine multiplikativ überlagerte Variable (mit Erwartungswert des Überlagerungsfaktors von Eins) ergibt sich ebenfalls Erwartungstreue. Für die Varianz gilt hingegen im Fall der multiplikativen Überlagerung:

$$(11) \qquad \mathrm{var}\left(X^{a(mult)}\right) = \sigma_x^2 + (\sigma_x^2 + \mu_x^2)\sigma_w^2.$$

Setzt man die beiden Formeln für die Varianz gleich und löst nach σ_v^2 auf, so ergibt sich:

$$(12) \qquad \sigma_v^2 = (\sigma_x^2 + \mu_x^2)\sigma_w^2 .$$

Das gleiche Ergebnis erhält man, wenn man die multiplikative Überlagerung direkt als additiven Fehler betrachtet. Es ergibt sich dann der folgende Zusammenhang:

$$(13) \qquad X + V = XW \text{ oder } V = X(W-1) .$$

Der Erwartungswert von V ist auch in diesem Fall wieder Null. Für die Varianz ergibt sich wiederum das in Gleichung (12) gefundene Ergebnis. Folglich kann auch im Fall einer multiplikativen stochastischen Überlagerung mit einem SIMEX-Schätzer, der für additive Überlagerungen programmiert wurde, eine Korrektur durchgeführt werden.

Die Ausführungen zeigen, dass die Konstruktion des SIMEX-Schätzers sehr einfach und anschaulich ist. Allerdings sind die Eigenschaften des Schätzers sehr komplex. Deshalb ist auch die Berechnung der Standardfehler und Teststatistiken schwierig. Carroll u. a.[73]) untersuchen die asymptotische Verteilung des SIMEX-Schätzers für parametrische Modelle. Unter der Annahme, dass die Variablen identisch und unabhängig verteilt sind, zeigen sie, dass der SIMEX-Schätzer asymptotisch normalverteilt ist und leiten einen Schätzer für dessen asymptotische Varianz-Kovarianzmatrix her. Stefanski und Cook[74]) leiten eine Methode der Varianzschätzung her, die angewendet werden kann, wenn die Varianz der Überlagerung bekannt ist oder so gut geschätzt werden kann, dass man diese Annahme treffen kann.

5 Mikroaggregation in linearen und nichtlinearen Modellen: Ein Überblick

5.1 Mikroaggregation im linearen Modell

Zunächst wird der Fall betrachtet, dass eine zufällige Mikroaggregation vorgenommen wird, bei der die Aggregationsmatrix $\mathbf{D}$ von den zu anonymisierenden Variablen unabhängig ist, oder eine abstandsorientierte Mikroaggregation, bei der die Aggregationsmatrix ausschließlich von den Regressoren abhängt und nicht von der abhängigen Variablen. Es lässt sich zeigen, dass in diesem Fall die Aggregationsmatrix symmetrisch idempotent ist.[75])[76]) Für diese gilt: $\mathbf{D} = \mathbf{I}_M \otimes \frac{1}{A}\mathbf{u}\mathbf{u}'$. Dabei ist M die Anzahl der Gruppen und A die Gruppengröße.

Wird lediglich die abhängige Variable mikroaggregiert, so ergibt sich für den Erwartungswert des OLS-Schätzers:[77])

$$(14) \qquad E\left(\hat{\beta}\right) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}\mathbf{X}\beta$$

und für seinen Wahrscheinlichkeitsgrenzwert:

$$(15) \qquad p\lim\hat{\beta} = \mathbf{Q}^{-1}p\lim\left(\frac{\mathbf{X}'\mathbf{D}\mathbf{X}}{n}\right)\beta .$$

Damit ist der Schätzer in diesem Fall weder erwartungstreu noch konsistent. Anders sieht es aus, wenn statt der abhängigen Variablen lediglich die Regressoren oder abhängige Variable und Regressoren gemeinsam mikroaggregiert werden. In diesem Fall ist der OLS-Schätzer erwartungstreu. Allerdings ist er nicht effizient. Lechner und Pohlmeier[78]) leiten her, dass der Effizienzverlust umso geringer ist, je kleiner die Gruppengröße bei der Mikroaggregation gewählt wird und je ähnlicher sich die in einer Gruppe zusammengefassten Werte sind. Schmid u. a.[79]) zeigen, dass der Effizienzverlust asymptotisch gegen Null geht.

Durch die Mikroaggregation wird auch der herkömmliche Schätzer für die Varianz des Fehlerterms und somit der Standardfehler des OLS-Schätzers verzerrt. Lechner und Pohlmeier leiten folgenden erwartungstreuen Schätzer für die Residuenvarianz im Fall gemeinsam mikroaggregierter Variablen her:[80])

$$(16) \qquad \hat{\sigma}_u^2 = \frac{\hat{\mathbf{u}}^{a\prime}\hat{\mathbf{u}}^a}{M-K}.$$

Dabei ist $\hat{\mathbf{u}}^a$ der Vektor der geschätzten Residuen, M ist die Anzahl der durch die Mikroaggregation gebildeten Gruppen und K die Anzahl der Merkmale.

Werden nun nicht mehr alle Merkmale gemeinsam mikroaggregiert, sondern nur teilweise, so kann man das lineare Regressionsmodell ganz allgemein wie folgt schreiben:

$$(17) \qquad \mathbf{D}_y\mathbf{y} = \beta_1\mathbf{D}_1\mathbf{X}_1 + \beta_2\mathbf{D}_2\mathbf{X}_2 + \mathbf{u} .$$

Hierzu werden die Regressoren in zwei Gruppen $\mathbf{X}_1$ und $\mathbf{X}_2$ aufgeteilt, die jeweils gemeinsam mikroaggregiert werden oder gegebenenfalls auch im Originalzustand verbleiben können.

Mit Hilfe dieser Darstellung kann man für folgende Fälle herleiten, dass der OLS-Schätzer weder erwartungstreu noch konsistent ist:[81])

73) Siehe Carroll, R./Küchenhoff, H./Lombard, F./Stefanski, L.: "Asymptotics for the SIMEX Estimator in Nonlinear Measurement Error Models", Journal of the American Statistical Association, Bd. 91(433), 1996, S. 242 ff.
74) Siehe Fußnote 72.
75) Siehe Fußnote 55.
76) Dies gilt nicht für den Fall der Bootstrap-Aggregation. Siehe dazu Rosemann, M., a. a. O., Fußnote 17, Kapitel 12.
77) Siehe Ronning, G. u. a., a. a. O., Fußnote 13, Kapitel 23, sowie Rosemann, M., a. a. O., Fußnote 17, Kapitel 17.
78) Siehe Fußnote 55.
79) Siehe Schmid, M./Schneeweiss, H./Küchenhoff, H.: "Consistent Estimation of a Simple Linear Model under Microaggregation", SFB Discussion Paper No. 415, 2005.
80) Siehe Lechner, S./Pohlmeier, W., a. a. O., Fußnote 55.
81) Siehe Fußnote 77.

(1) $\mathbf{D}_y \neq \mathbf{D}_1 = \mathbf{D}_2$: Die abhängige Variable wird getrennt von den gemeinsam mikroaggregierten Regressoren mikroaggregiert.

(2) $\mathbf{D}_y \neq \mathbf{D}_1 \neq \mathbf{D}_2$: Die abhängige Variable wird getrennt von den teilweise gemeinsam mikroaggregierten Regressoren mikroaggregiert.

(3) $\mathbf{D}_y = \mathbf{D}_1 = \mathbf{I}_n \neq \mathbf{D}_2$: Lediglich ein Teil der Regressoren wird gemeinsam mikroaggregiert.

(4) $\mathbf{D}_y = \mathbf{D}_1 \neq \mathbf{D}_2 = \mathbf{I}_n$: Die abhängige Variable und ein Teil der Regressoren werden gemeinsam mikroaggregiert.

(5) $\mathbf{D}_y \neq \mathbf{D}_1 \neq \mathbf{D}_2 = \mathbf{I}_n$: Die abhängige Variable und ein Teil der Regressoren werden getrennt voneinander mikroaggregiert.

Zusammengefasst ergibt sich damit bei den bisher behandelten Varianten nur in zwei Fällen ein erwartungstreuer beziehungsweise konsistenter KQ-Schätzer: zum einen, wenn ausschließlich die Regressoren (und zwar alle gemeinsam) mikroaggregiert werden, zum anderen, wenn abhängige Variable und Regressoren gemeinsam mikroaggregiert werden. Dieses Ergebnis gilt sowohl für die zufällige Mikroaggregation als auch für eine deterministische Mikroaggregation, bei der die abstandsorientierte Gruppenbildung nur von den Regressoren, nicht aber von der abhängigen Variablen abhängt.[82])

Diese Ergebnisse gelten nicht, falls das Gewichtungsschema der Aggregation von der abhängigen Variablen abhängt [$\mathbf{D} = \mathbf{D}(\mathbf{y})$ oder $\mathbf{D} = \mathbf{D}(\mathbf{X},\mathbf{y})$]. Der Hauptunterschied zu der bisher betrachteten Situation besteht darin, dass die Aggregationsmatrix, beziehungsweise das zur Gruppenbildung verwendete Abstandsmaß, nun vom Fehlerterm des Modells abhängt.[83]) Schmid u. a. (2005) und Schmid (2006) leiten den Wahrscheinlichkeitsgrenzwert des Schätzers für diese Fälle ab und bestimmen daraus die entsprechenden Korrekturschätzer.[84])

Wesentlich ist die Betrachtung der abstandsorientierten getrennten Mikroaggregation, bei der jede Variable getrennt sortiert und anschließend mikroaggregiert wird. Für diese spezielle Variante der Mikroaggregation, die auch als „Individual Ranking“ bezeichnet wird, zeigt Schmid, dass der OLS-Schätzer konsistent ist.[85])

Abschließend wird noch die in Kapitel 3 ebenfalls eingeführte Bootstrap-Mikroaggregation untersucht. Wird ausschließlich die abhängige Variable mikroaggregiert, so kann aus den Herleitungen für die gewöhnliche Mikroaggregation auch für die Bootstrap-Aggregation eine Verzerrung des Schätzers gefolgert werden.

Werden ausschließlich die Regressoren mikroaggregiert, so ergibt sich für den Erwartungswert des Schätzers:

$$E\left(\hat{\beta}\right) = \left(\mathbf{X}'\mathbf{D}_{BS}'\mathbf{D}_{BS}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{D}_{BS}'\mathbf{X}\beta. \tag{18}$$

Der Schätzer ist in diesem Fall nicht erwartungstreu, weil die stochastische Aggregationsmatrix $\mathbf{D}_{BS}$ nicht symmetrisch idempotent ist. Wird hingegen auch die abhängige Variable in diese Form der Mikroaggregation einbezogen, ist der Schätzer wiederum erwartungstreu.[86])

5.2 Mikroaggregation in nichtlinearen Modellen

Bei der Mikroaggregation handelt es sich um eine lineare Transformation. Einem nichtlinearen Modell liegt ein nichtlinearer Zusammenhang und somit eine nichtlineare Transformation zugrunde. Lineare und nichtlineare Transformationen sind jedoch nicht vertauschbar. Damit werden nichtlineare Zusammenhänge durch die Mikroaggregation zerstört. Folglich werden auch die Schätzer in nichtlinearen Modellen durch die Mikroaggregation grundsätzlich verzerrt. Umfang und Art der Verzerrung sind dabei stark von der Art der verwendeten Mikroaggregation sowie der Art des nichtlinearen Zusammenhangs und der konkreten Modellspezifikation abhängig.[87])

6 Praxisbeispiele mit Daten der Kostenstrukturerhebung im Verarbeitenden Gewerbe und des IAB-Betriebspanels

6.1 Schätzung einer linearisierten Cobb-Douglas-Produktionsfunktion mit Daten der Kostenstrukturerhebung im Verarbeitenden Gewerbe

In Anlehnung an das Vorgehen von Fritsch und Stephan[88]) werden die Produktionselastizitäten einer Cobb-Douglas-Produktionsfunktion für das Verarbeitende Gewerbe einschließlich Bergbau sowohl mit den Originaldaten der Kostenstrukturerhebung im Verarbeitenden Gewerbe (KSE) als auch mit anonymisierten KSE-Daten geschätzt. Im Unterschied zu Fritsch und Stephan stehen lediglich Querschnittsdaten für das Jahr 1999 zur Verfügung.[89]) Die Cobb-Douglas-Produktionsfunktion hat folgende Gestalt:

82) Im verallgemeinerten Modell ergibt sich sogar nur in dem Fall Erwartungstreue, in dem alle Variablen gemeinsam mikroaggregiert werden. Siehe dazu Fußnote 77.
83) Siehe Fußnote 79.
84) Siehe Fußnote 79 sowie Fußnote 51.
85) Siehe Schmid, M., a. a. O., Fußnote 51.
86) Siehe Fußnote 77.
87) Siehe Fußnote 21, Kapitel 23, sowie Fußnote 77.
88) Siehe Fritsch, M./Stephan, A.: „Die Heterogenität der technischen Effizienz innerhalb von Wirtschaftszweigen – Auswertungen auf Grundlage der Kostenstrukturstatistik des Statistischen Bundesamtes“ in Pohl, R./Fischer, J./Rockmann, U./Semlinger, K. (Hrsg.): „Analysen zur regionalen Industrieentwicklung – Sonderauswertungen einzelbetrieblicher Daten der Amtlichen Statistik“, Berlin 2003, S. 143 ff.
89) Eine ausführliche Datensatzbeschreibung findet sich in Ronning, G. u. a., a. a. O., Fußnote 13, Kapitel 9, und Rosemann, M., a. a. O., Fußnote 17, Kapitel 11.

$$(19) \qquad Y = A\prod_{k=1}^{K} X_k^{\beta_k}.$$

Dabei ist A der konstante Technologieparameter, β_k sind die Produktionselastizitäten, X_k die Inputfaktoren und Y der Output. Der Output wird wie bei Fritsch und Stephan als Bruttoproduktionswert vermindert um den Umsatz aus sonstiger Tätigkeit definiert. Als Inputfaktoren werden in Anlehnung an Fritsch und Stephan der Materialeinsatz, die Personalkosten, die Kosten für externe Dienstleistungen, die Kapitalkosten und die Sonstigen Kosten verwendet.[90])

Durch Bildung des Logarithmus auf beiden Seiten der Gleichung (19) kann die Funktion linearisiert werden. Man schätzt daher das folgende lineare Modell:

$$(20) \qquad \log(Y) = C + \sum_{k=1}^{K} \beta_k \log(X_k).$$

Grundsätzlich werden Unternehmen ausgeschlossen, bei denen ein Inputfaktor oder der Output den Wert Null aufweist. Um zu vermeiden, dass Unternehmen mit extremen Merkmalsausprägungen einen zu großen Einfluss auf die Schätzergebnisse haben, wird alternativ zur Berücksichtigung aller Unternehmen im Datensatz ein Szenario geschätzt, bei dem der Datensatz vorher wie folgt „bereinigt" wird: Unternehmen, bei denen der Produktionsanteil eines Inputfaktors weniger als das 1-Prozent-Quantil bzw. mehr als das 99-Prozent-Quantil der Verteilung der Produktionsanteile über alle Unternehmen beträgt, werden gemäß dem Vorgehen von Fritsch und Stephan von den Berechnungen ausgeschlossen. Für beide Szenarien sind die Schätzergebnisse mit den Originaldaten der Kostenstrukturerhebung im Verarbeitenden Gewerbe in Tabelle 1 dargestellt. Die Standardfehler werden jeweils robust geschätzt.

Tabelle 1: Cobb-Douglas-Produktionsfunktion – Schätzergebnisse für Originaldaten der Kostenstrukturerhebung im Verarbeitenden Gewerbe 1999

Koeffizienten	Daten nicht bereinigt	Daten bereinigt
Materialeinsatz	0,414	0,435
(t-Werte)	(79,02)	(149,95)
Personalkosten	0,339	0,322
(t-Werte)	(58,87)	(100,69)
Externe Dienstleistungen	0,058	0,052
(t-Werte)	(29,27)	(39,31)
Sonstige Kosten	0,114	0,105
(t-Werte)	(35,01)	(46,91)
Kapitalkosten	0,055	0,07
(t-Werte)	(16,14)	(32,44)
Konstante	1,805	1,717
(t-Werte)	(66,42)	(97,88)
Anzahl der Beobachtungen	16 343	15 017
R^2	0,977	0,988

Quelle: Eigene Berechnungen.

Um die Wirkung der in den Kapiteln 4 und 5 beschriebenen Anonymisierungsverfahren zu veranschaulichen, werden anschließend die Ergebnisse von Modellschätzungen dargestellt, bei denen die in das Modell eingehenden Variablen (der logarithmierte Output sowie die logarithmierten Inputs) stochastisch überlagert bzw. mikroaggregiert wurden.[91])

Die Ergebnisse für die stochastischen Überlagerungen beziehen sich auf den Fall unbereinigter Daten. In jeder Variante werden alle in das Modell einbezogenen Merkmale überlagert. Folgende Varianten der stochastischen Überlagerung werden betrachtet:

- Additive Überlagerung der sechs Variablen mit voneinander unabhängigen standardnormalverteilten Zufallsfehlern;
- Additive Überlagerung der sechs Variablen mit Zufallsfehlern, deren Varianz-Kovarianzmatrix proportional (Faktor $d = 0{,}1$) zur Varianz-Kovarianzmatrix der Originalvariablen ist;
- Additive Überlagerung nach dem Verfahren von Kim (Faktor $d = 0{,}1$);
- Multiplikative Überlagerung der sechs Variablen mit einem für jedes Unternehmen konstanten Faktor, dabei entstammen die Faktoren einer Gleichverteilung mit dem Intervall (0,8; 1,2);
- Multiplikative Überlagerung der sechs Variablen mit unterschiedlichen Faktoren, die ebenfalls einer Gleichverteilung mit dem Intervall (0,8; 1,2) entstammen;
- Multiplikative Überlagerung der sechs Variablen mit Faktoren aus einer zweigipfligen Mischungsverteilung nach dem Verfahren von Höhne ($f = 0{,}11$, $s = 0{,}03$).

Die Untersuchung erfolgt im Rahmen von Monte-Carlo-Simulationen mit 1 000 Replikationen, das heißt die stochastische Überlagerung und die anschließende Modellschätzung wird für jede Anonymisierungsvariante 1 000-mal wiederholt. Die relativen Abweichungen der Koeffizientenschätzer sind in den Tabellen 2 und 3 dargestellt.[92])

Tabelle 2: Linearisierte Cobb-Douglas-Produktionsfunktion – Relative Abweichungen der Schätzergebnisse für additiv überlagerte Merkmale im Vergleich zu den Ergebnissen mit Originaldaten, 1 000 Replikationen

Prozent

Koeffizienten	Additive Überlagerung standardnormalverteilt		Additive Überlagerung, normalverteilt, VCV-Matrix der Überlagerungen proportional zur VCV-Matrix der Originalvariablen	Additive Überlagerung Kim-Verfahren
	keine Korrektur	IV-Schätzung		
Materialeinsatz	35,35	0,03	0,09	0,10
Personalkosten	47,83	0,13	0,10	0,11
Externe Dienstleistungen	99,35	0,01	0,00	0,01
Sonstige Kosten	57,25	0,33	0,34	0,31
Kapitalkosten	146,63	0,71	0,06	0,07
Konstante	115,45	0,04	0,01	0,01

Quelle: Eigene Berechnungen.

90) Zu den Details siehe Rosemann, M., a. a. O., Fußnote 17, Kapitel 15.

91) Zu dem in der Praxis relevanteren Fall anonymisierter Ausgangsvariablen siehe Ronning, G. u. a., a. a. O., Fußnote 13, Kapitel 22 und 23, sowie Rosemann, M., a. a. O., Fußnote 17, und Fußnote 77.

92) Ausführliche Ergebnisse finden sich in Ronning, G. u. a., a. a. O., Fußnote 13, und Rosemann, M., a. a. O., Fußnote 17.

Tabelle 3: Linearisierte Cobb-Douglas-Produktionsfunktion – Relative Abweichungen der Schätzergebnisse für multiplikativ überlagerte Merkmale im Vergleich zu den Ergebnissen mit Originaldaten, 1 000 Replikationen
Prozent

Koeffizienten	Multiplikative Überlagerung, konstante Faktoren, Gleichverteilung (0,8; 1,2)		Multiplikative Überlagerung, unterschiedliche Faktoren, Gleichverteilung (0,8; 1,2)		Multiplikative Überlagerung, Mischungsverteilung nach dem Verfahren von Höhne ($f = 0{,}11$; $s = 0{,}03$)	
	keine Korrektur	IV-Schätzung	keine Korrektur	IV-Schätzung	keine Korrektur	IV-Schätzung
Materialeinsatz	2,43	0,10	58,73	0,04	10,92	0,10
Personalkosten	35,67	0,03	64,16	0,20	2,27	0,19
Externe Dienstleistungen	36,63	0,18	151,67	1,77	6,80	0,31
Sonstige Kosten	25,39	0,49	48,86	1,80	27,37	0,63
Kapitalkosten	15,66	0,19	157,56	0,93	118,63	0,47
Konstante	68,99	0,15	230,58	0,63	44,49	0,29

Quelle: Eigene Berechnungen.

Die additive Überlagerung mit standardnormalverteilten Zufallsfehlern führt zu verzerrten Schätzern, die sich beispielsweise durch eine Instrumentvariablen-Schätzung wieder korrigieren lassen.[93]) Die additive Überlagerung mit Zufallsfehlern, deren Varianz-Kovarianzmatrix proportional zur Varianz-Kovarianzmatrix der Originalvariablen gewählt wird, und das Kim-Verfahren erhalten die Koeffizientenschätzer. Auch die Schätzer im Fall einer multiplikativen stochastischen Überlagerung sind verzerrt, unabhängig davon, ob mit einem konstanten Faktor überlagert wird oder nicht. Während sich bei Anwendung des Verfahrens von Höhne und bei Verwendung konstanter Überlagerungsfaktoren mit dem IV-Schätzer hier eine Korrektur vornehmen lässt, gelingt dies bei unterschiedlichen Faktoren zwar im Mittel über alle 1 000 Simulationsläufe, jedoch in vielen einzelnen Simulationsläufen nicht. Hier scheint offenbar die gewählte Überlagerungsvarianz zu hoch zu sein.

Bei der Untersuchung der Mikroaggregationsverfahren werden hingegen die bereits um Ausreißer bereinigten Daten

Tabelle 4: Linearisierte Cobb-Douglas-Produktionsfunktion – Relative Abweichungen der Schätzergebnisse für mikroaggreggierte Merkmale im Vergleich zu den Ergebnissen mit Originaldaten, 1 000 Replikationen
Prozent

Koeffizienten	Mikroaggregation			
	getrennt abstandsorientiert	gemeinsam abstandsorientiert	gemeinsam zufällig	gemeinsam Bootstrap
Materialeinsatz	0	0	0,69	0
Personalkosten	0	0	1,24	0,62
Externe Dienstleistungen	0	0	1,92	0
Sonstige Kosten	0	0	1,90	0,95
Kapitalkosten	0	2,86	4,29	2,86
Konstante	0	0,82	0,52	0,12

Quelle: Eigene Berechnungen.

verwendet, weshalb die in Tabelle 4 dargestellten Ergebnisse auch nicht direkt mit denen in den Tabellen 2 und 3 vergleichbar sind. Im Einzelnen werden folgende Varianten der Mikroaggregation untersucht:

- Getrennte abstandsorientierte Mikroaggregation (Individual Ranking)
- Gemeinsame abstandsorientierte Mikroaggregation (Euklidische Distanz über alle Merkmale der Kostenstrukturerhebung im Verarbeitenden Gewerbe)[94])
- Gemeinsame zufällige Mikroaggregation
- Bootstrap-Mikroaggregation für alle Merkmale gemeinsam[95])

Wie in Kapitel 5 hergeleitet wurde, führt die getrennte abstandsorientierte Mikroaggregation zu unverzerrten Schätzern. Dies müsste nach den theoretischen Ableitungen auch für die gemeinsame zufällige Mikroaggregation und die Bootstrap-Aggregation gelten. Hier zeigen die Schätzergebnisse geringfügige Abweichungen. Zuletzt führt die gemeinsame abstandsorientierte Mikroaggregation dann zu einem verzerrten Schätzer, wenn die Aggregationsmatrix (auch) von der abhängigen Variablen abhängt, was hier der Fall ist. Allerdings ist der Einfluss der abhängigen Variablen in diesem Fall sehr gering, weil alle 30 metrischen Merkmale der Kostenstrukturerhebung im Verarbeitenden Gewerbe zur Berechnung des Abstandsmaßes herangezogen wurden. Zudem geht die Verzerrung gegen Null, falls abhängige und erklärende Variable hoch korreliert sind, was in diesem Fall ebenfalls zutrifft. Beides erklärt daher die ebenfalls nur geringfügigen Abweichungen der Schätzkoeffizienten bei der gemeinsamen abstandsorientierten Mikroaggregation.

6.2 Schätzung eines binären Probit-Modells zur Erklärung der Tarifbindung mit Daten des IAB-Betriebspanels

Die Entscheidung eines Betriebes, sich der Tarifbindung zu unterziehen, wird in dieser Studie unter Verwendung einer einfachen Maximum-Likelihood-Probit-Schätzung untersucht. Es wird angenommen, dass der latente Nutzen des Betriebes, einen Tarifvertrag zu akzeptieren, durch die lineare Kombination der beobachtbaren Determinanten und eine i. i. d. (independent and identically distributed) verteilte Störgröße (für die unbeobachtete Heterogenität) beschrieben werden kann. Die wichtigsten Determinanten der Tarifbindung sind die Größe des Betriebes und die Branchenzugehörigkeit. Weitere Determinanten sind zum Beispiel das

93) Als Instrumente werden hier in gleicher Weise anonymisierte Daten verwendet. Es wird also davon ausgegangen, dass der Nutzer zwei anonymisierte Datensätze erhält.
94) Die Anonymisierung bei dieser Variante wurde von Jörg Höhne (Statistik Berlin-Brandenburg) vorgenommen.
95) Auch bei den beiden stochastischen Mikroaggregationsvarianten wird jeweils nur eine Lösung getestet.

Betriebsalter, die Existenz eines Betriebsrats oder der Anteil von qualifizierten Angestellten.[96])

Für die Replikationsstudie wird ein „Basismodell" geschätzt. Dabei dienen die Betriebsgröße (gemessen durch den natürlichen Logarithmus der Beschäftigung) und ein Satz von Dummy-Variablen für verschiedene Branchen als erklärende Variable.[97]) Die abhängige Variable nimmt den Wert Eins an, falls der Betrieb tarifgebunden ist, und den Wert Null, falls der Betrieb nicht tarifgebunden ist. Für die Schätzungen werden die Daten des IAB-Betriebspanels für Baden-Württemberg verwendet, das für die hier betrachteten Branchen insgesamt 1 201 Betriebe umfasst. Die Schätzergebnisse mit den Originaldaten sind in Tabelle 5 dargestellt.

Tabelle 5: ML-Probitschätzung zur Erklärung der Tarifbindung mit Originaldaten des IAB-Betriebspanels Baden-Württemberg 2001

Koeffizienten	Werte
Log. Beschäftigung	0,301
(t-Wert)	(12,24)
Bau	0,748
(t-Wert)	(4,46)
Handel	0,379
(t-Wert)	(2,88)
Dienstleistungen	0,040
(t-Wert)	(0,40)
Verwaltung	0,796
(t-Wert)	(4,60)
Konstante	–0,988
(t-Wert)	(–7,46)
Beobachtungen	1 201
Pseudo R^2	0,140
LR-test	222,78
Log. likelihood	–686,46

Quelle: Strotmann (2004); siehe Fußnote 96 im Text.

Anonymisiert wird jeweils die logarithmierte Beschäftigung. Folgende Anonymisierungsmethoden werden getestet:

- Additive stochastische Überlagerung mit einer Normalverteilung. Die Varianz beträgt das 0,1-Fache der Ausgangsvarianz der logarithmierten Beschäftigung.
- Multiplikative Überlagerung mit einer Gleichverteilung im Intervall (0,1; 1,5)
- Multiplikative Überlagerung mit einer Mischungsverteilung nach dem Verfahren von Höhne ($f = 0{,}11$; $s = 0{,}03$)
- Abstandsorientierte Mikroaggregation
- Zufällige Mikroaggregation

Bei den stochastischen Überlagerungen und der stochastischen Mikroaggregation werden 1 000 Replikationen durchgeführt. Betrachtet wird der Durchschnitt aus allen Simulationsläufen. Die relativen Abweichungen der Schätzkoeffizienten sind in den Tabellen 6 und 7 dargestellt.

Beim Verfahren der stochastischen Überlagerung tritt in allen Fällen eine Verzerrung der Schätzer auf, die durch die Anwendung des in Kapitel 4 vorgestellten SIMEX-Schätzers im Wesentlichen korrigiert werden kann. Eine Ausnahme stellt bei der additiven Überlagerung der Koeffizient für den allerdings ohnehin insignifikanten Dienstleistungssektor dar. Bei der Anwendung des SIMEX-Schätzers wird die Varianz der Überlagerung geschätzt. Der Fall der multiplikativen Überlagerung wird gemäß dem in Kapitel 4 beschriebenen Vorgehen als additive Überlagerung interpretiert.

Aus den in Tabelle 7 dargestellten Ergebnissen erkennt man, dass sich im Fall der abstandsorientierten Mikroaggregation annähernd die gleichen Ergebnisse erzielen lassen wie mit den Originaldaten. Demgegenüber ergeben sich durch die zufällige Mikroaggregation bei der Probit-Schätzung starke Abweichungen der Schätzkoeffizienten gegenüber den mit den Originaldaten erzielten Ergebnissen. Die Unterscheidung zwischen getrennter und gemeinsamer abstandsorientierter Mikroaggregation ist im Fall nur einer

Tabelle 7: ML-Probitschätzung zur Erklärung der Tarifbindung – Relative Abweichungen der Schätzergebnisse für mikroaggregierte logarithmierte Beschäftigung im Vergleich zu den Ergebnissen mit Originaldaten

Koeffizienten	Abstandsorientierte Mikroaggregation	Zufällige Mikroaggregation
Log. Beschäftigung	0,09	20,12
Bau	0,11	67,41
Handel	0,34	122,80
Dienstleistungen	0,85	878,48
Verwaltung	0,14	13,79
Konstante	0,11	44,79

Quelle: Eigene Berechnungen.

Tabelle 6: ML-Probitschätzung zur Erklärung der Tarifbindung – Relative Abweichungen der Schätzergebnisse für stochastisch überlagerte logarithmierte Beschäftigung im Vergleich zu den Ergebnissen mit Originaldaten

Prozent

Koeffizienten	Additive Überlagerung (d = 0,1)		Multiplikative Überlagerung, Gleichverteilung (0,5; 1,5)		Multiplikative Überlagerung, Mischungsverteilung nach dem Verfahren von Höhne ($f = 0{,}11$; $s = 0{,}03$)	
	keine Korrektur	SIMEX-Korrektur	keine Korrektur	SIMEX-Korrektur	keine Korrektur	SIMEX-Korrektur
Log. Beschäftigung	35,13	3,13	33,14	0,58	6,76	0,66
Bau	27,84	2,44	27,59	0,49	5,79	0,32
Handel	51,28	4,47	49,93	0,61	10,45	0,81
Dienstleistungen	371,51	32,21	358,02	0,07	74,35	6,68
Verwaltung	5,21	0,51	5,21	0,46	1,03	0,14
Konstante	49,85	4,39	47,83	0,03	0,86	0,85

Quelle: Eigene Berechnungen.

96) Siehe Strotmann, H.: „Tarifbindung in Baden-Württemberg im Jahr 2000 – ist der Flächentarifvertrag ein Auslaufmodell?", IAW-Mitteilungen, Bd. 1/2002, S. 4 ff.

97) Siehe Strotmann, H.: "The Impact of Anonymisation on Binary Choice Models – Empirical Evidence from Monte Carlo Simulations using the IAB Establishment Panel Baden-Wuerttemberg", Beitrag zum Workshop "Econometric Analysis of Anonymised Firm Data", Tübingen, März 2004.

Tabelle 8: Additive stochastische Überlagerung im linearen Modell

Art der Überlagerung	Nur abhängige Variable stochastisch überlagert	Alle Variablen stochastisch überlagert	Nur Regressoren oder Teil der Regressoren oder Teil der Regressoren und abhängige Variable überlagert
Einfache additive Überlagerung	erwartungstreu	nicht (asymptotisch) erwartungstreu, korrigierbar (z. B. Fuller 1987, IV)	nicht (asymptotisch) erwartungstreu, korrigierbar (z. B. Fuller 1987, IV)
VCV-Matrix der Überlagerungen proportional zur VCV-Matrix der Originaldaten	erwartungstreu	asymptotisch erwartungstreu	nicht (asymptotisch) erwartungstreu, korrigierbar (z. B. Fuller 1987, IV)
Überlagerung mit proportionaler VCV-Matrix und Transformation zum Erhalt der ersten und zweiten Momente (Verfahren von Kim)	erwartungstreu	asymptotisch erwartungstreu, t-Werte erhalten	nicht (asymptotisch) erwartungstreu, korrigierbar (z. B. Fuller 1987, IV)
Erste und zweite Momente werden gegebenenfalls auch für nicht überlagerte Merkmale erhalten	erwartungstreu	asymptotisch erwartungstreu, t-Werte erhalten	asymptotisch erwartungstreu, t-Werte erhalten

Quelle: Eigene Darstellung.

metrischen Regressorvariablen natürlich nicht möglich. Bei mehreren metrischen Regressoren würde sich jedoch im Fall der gemeinsamen abstandsorientierten Mikroaggregation aus den in Kapitel 5 genannten Gründen eine Verzerrung der Schätzkoeffizienten ergeben.

7 Zusammenfassung und Ausblick

Die aufgezeigten Auswirkungen von stochastischen Überlagerungen und Mikroaggregationsverfahren im linearen Modell sind in den Tabellen 8 bis 10 zusammenfassend dargestellt. Für nichtlineare Modelle können folgende Ergebnisse festgehalten werden:

- Additive und multiplikative stochastische Überlagerungen führen zu einer Verzerrung der Koeffizientenschätzer in nichtlinearen Modellen, die jedoch durch die Anwendung von Korrekturschätzern korrigiert werden können. In der Anwendung besonders einfach ist die SIMEX-Methode.

- Mikroaggregationsverfahren, bei denen die Merkmale gemeinsam mikroaggregiert werden, führen zu verzerrten Koeffizientenschätzern. Bei der getrennten abstandsorientierten Mikroaggregation können die Koeffizienten jedoch auch in nichtlinearen Modellen gut geschätzt werden.

Da ein Scientific-Use-File möglichst viele Nutzungspotenziale eröffnen soll, bieten sich stochastische Überlagerungen und die getrennte abstandsorientierte Mikroaggregation als datenverändernde Anonymisierungsverfahren an. Die Mikroaggregation weist dabei den Vorteil auf, dass die Anwendung eines Korrekturverfahrens nicht erforderlich ist. Die in diesem Beitrag vorgestellten Varianten der multiplikativen stochastischen Überlagerung weisen gegenüber additiven Überlagerungen aus der Sicht der Datennutzer den Vorteil auf, dass sie die Vorzeichen ebenso erhalten wie die (strukturellen) Nullen. Sowohl die abstandsorientierte getrennte Mikroaggregation als auch multiplikative stochastische Überlagerungen erlauben auch eine flexible Ergänzung von Datensätzen durch zusätzliche von den Nutzern nachgefragte nichtlineare Transformationen aus den Ausgangsvariablen.

In jedem Fall stellt die Anwendung datenverändernder Anonymisierungsverfahren zur Anonymisierung von Mikrodaten einen Eingriff in das Analysepotenzial dieser Daten dar. Ob dieser Eingriff notwendig und gerechtfertigt ist, hängt vom konkreten Fall ab. Allerdings soll abschließend nochmals darauf hingewiesen werden, dass die mit den als brauchbar bewerteten datenverändernden Anonymisierungsverfahren verbundenen Einschränkungen des Analysepotenzials in vielen Fällen für die Nutzer weitaus akzeptabler sein dürften als die Einschränkung der Analysemöglichkeiten durch informationsreduzierende Maßnahmen, wie das Entfernen von Merkmalen, das Umwandeln von metrischen Merkmalen in kategoriale oder die weitgehende Zusammenfassung bereits bestehender Kategorien. Insofern bietet sich mit dem Einsatz der getrennten abstandsorientierten Mikroaggregation oder multiplikativer stochastischer Überlagerungen gerade auch als Ergänzung zu informationsreduzierenden Maßnahmen ein vielversprechender Weg zur

Tabelle 9: Multiplikative stochastische Überlagerung im linearen Modell

Art der Überlagerung	Nur abhängige Variable stochastisch überlagert	Alle Variablen stochastisch überlagert	Nur Regressoren oder Teil der Regressoren oder Teil der Regressoren und abhängige Variable überlagert
Multiplikative Überlagerung mit Erwartungswert Eins, kein konstanter Faktor	erwartungstreu	nicht (asymptotisch) erwartungstreu, korrigierbar (z. B. Hwang 1986, IV)	nicht (asymptotisch) erwartungstreu, korrigierbar (z. B. Hwang 1986, IV)
Multiplikative Überlagerung mit Erwartungswert Eins, konstanter Faktor .	erwartungstreu	unter bestimmten Bedingungen konsistent (kein Absolutglied, Mittelwerte aller Regressoren gleich Null), ansonsten korrigierbar (z. B. Hwang 1986, IV)	nicht (asymptotisch) erwartungstreu, korrigierbar (z. B. Hwang 1986, IV)

Quelle: Eigene Darstellung.

Tabelle 10: Mikroaggregation im linearen Modell

Art der Mikroaggregation	Alle Variablen mikroaggregiert	Nur abhängige Variable mikroaggregiert	Nur alle Regressoren mikroaggregiert	Nur Teil der Regressoren oder Teil der Regressoren und abhängige Variable mikroaggregiert
Getrennte abstandsorientierte Mikroaggregation	asymptotisch erwartungstreu	asymptotisch erwartungstreu	asymptotisch erwartungstreu	asymptotisch erwartungstreu
Getrennte zufällige Mikroaggregation	nicht (asymptotisch) erwartungstreu	nicht (asymptotisch) erwartungstreu	nicht (asymptotisch) erwartungstreu, Ausnahme: nur ein Regressor	nicht (asymptotisch) erwartungstreu
Gemeinsame Mikroaggregation zufällig oder abstandsorientiert nach den Regressoren	erwartungstreu	nicht (asymptotisch) erwartungstreu	erwartungstreu	nicht (asymptotisch) erwartungstreu
Gemeinsame abstandsorientierte Mikroaggregation nach der abhängigen Variablen	nicht (asymptotisch) erwartungstreu, korrigierbar (Schmid 2006)	asymptotisch erwartungstreu	nicht (asymptotisch) erwartungstreu, korrigierbar (Schmid 2006)	nicht (asymptotisch) erwartungstreu
Gemeinsame abstandsorientierte Mikroaggregation nach abhängiger Variablen und Regressoren	nicht (asymptotisch) erwartungstreu, korrigierbar (Schmid 2006)	nicht (asymptotisch) erwartungstreu	nicht (asymptotisch) erwartungstreu, korrigierbar (Schmid 2006)	nicht (asymptotisch) erwartungstreu
Bootstrap-Mikroaggregation	erwartungstreu	nicht (asymptotisch) erwartungstreu	nicht (asymptotisch) erwartungstreu	nicht (asymptotisch) erwartungstreu

Quelle: Eigene Darstellung.

Erstellung von Scientific-Use-Files für wirtschaftsstatistische Mikrodaten an.

Die bisherigen Untersuchungen zur Wirkungsweise unterschiedlicher datenverändernder Anonymisierungsverfahren in linearen und nichtlinearen Modellen beschränken sich auf Querschnittsdaten. Allerdings stehen Paneldaten immer stärker im Zentrum des wirtschaftswissenschaftlichen Interesses. Aus diesem Grund werden die Untersuchungen zur Wirkung datenverändernder Verfahren im Rahmen eines Forschungsprojekts der Forschungsdatenzentren der Statistischen Ämter, des Forschungsdatenzentrums der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- und Berufsforschung sowie des Instituts für Angewandte Wirtschaftsforschung auf panelökonometrische Methoden ausgeweitet.

DESTATIS
wissen.nutzen.

Auszug aus Wirtschaft und Statistik

Herausgeber: Statistisches Bundesamt, Wiesbaden

Schriftleitung: N. N.
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
- www.destatis.de/kontakt