Forschungsdatenzentrum
der Bundesagentur für Arbeit
im Institut für Arbeitsmarkt-
und Berufsforschung

FDZ

FDZ-Methodenreport

01/2011
DE

Methodische Aspekte zu Arbeitsmarktdaten

Methodenreport: Synthetische Scientific-Use-Files der Welle 2007 des IAB-Betriebspanels

Jörg Drechsler

Bundesagentur für Arbeit

Methodenreport: Synthetische Scientific-Use-Files der Welle 2007 des IAB-Betriebspanels

Jörg Drechsler (Institut für Arbeitsmarkt und Berufsforschung)

Die FDZ-Methodenreporte befassen sich mit den methodischen Aspekten der Daten des FDZ und helfen somit Nutzerinnen und Nutzern bei der Analyse der Daten. Nutzerinnen und Nutzer können hierzu in dieser Reihe zitationsfähig publizieren und stellen sich der öffentlichen Diskussion.

FDZ-Methodenreporte (FDZ method reports) deal with methodical aspects of FDZ data and help users in the analysis of these data. In addition, users can publish their results in a citable manner and present them for public discussion.

Inhaltsverzeichnis

Zusammenfassung

Die Bereitstellung von Scientific-Use-Files für Betriebsdaten stellt für die Forschungsdatenzentren eine besondere Herausforderung dar. Aufgrund der kleineren Grundgesamtheiten, dem hohen Auswahlsatz und der oft extrem schiefen Verteilung einzelner Variablen ist eine Reidentifikation einzelner Befragungsteilnehmer wesentlich leichter möglich als beispielsweise bei Haushaltsbefragungen. Einfache Maßnahmen wie Vergröberungen bei einzelnen kategorialen Variablen sind daher nicht ausreichend, um den Datenschutz zu gewährleisten. Bei der Erzeugung synthetischer Datensätze wird versucht, ein möglichst exaktes Abbild der Originaldaten zu erzeugen, wobei sensible Merkmale und Merkmale, die zu einer Reidentifikation führen könnten durch mehrfach imputierte Werte ersetzt werden.
Neben einer Einführung in das Verfahren bietet dieser Methodenreport hilfreiche Hinweise, die es bei der Nutzung der synthetischen Datensätze zu beachten gilt. Zudem wird erklärt, wie der Datennutzer vorgehen muss, um mit den synthetischen Datensätzen valide Ergebnisse zu erhalten. Abschließend zeigen erste Analyseergebnisse das Potenzial aber auch die Grenzen der erzeugten Datensätze auf.

Abstract

Providing scientific use files for business surveys is a difficult task. Due to smaller populations, higher sampling rates, and skewed distributions disclosure risks are much higher than for household surveys. Simple measures like coarsening are not sufficient to protect the data. The aim of generating synthetic datasets is to release data that provide a high level of data utility while guaranteeing the confidentiality of the survey respondent. To achieve this, sensitive variables and variables that could be used for re-identification purposes are replaced with multiple imputations.
This report gives a short introduction to the topic and discusses some aspects that analysts should keep in mind when using the synthetic datasets. Furthermore, the report describes how valid inferences can be obtained based on the synthetic datasets and provides some first data utility evaluations that indicate the potentials but also the limits of the generated datasets.

Keywords: Anonymisierung, multiple Imputation, synthetische Datensätze, IAB-Betriebspanel.

Danksagung: Mein Dank gilt insbesondere Stefan Bender, der dieses Projekt ermöglicht hat. Außerdem danke ich Hans Kiesl für zahlreiche hilfreiche Diskussionen und Anmerkungen. John Abowd hat mir die Arbeit für Kapitel 4 dieses Reports sehr erleichtert indem er mir erlaubte, hemmungslos aus seinem technical report (Abowd/Stinson/Benedetto, 2006) zu kopieren. Zuletzt möchte ich Daniela Hochfellner danken, die mir bei der Formatierung und Überarbeitung geholfen hat.

1 Hintergrund

Neben rechtlichen Vorgaben und moralisch-ethischen Überlegungen hat das IAB auch aus Datenqualitätsgründen ein hohes Interesse, den zugesicherten Datenschutz zu garantieren. Die Betriebe, die sich oft schon seit vielen Jahren an der Erhebung des IAB Betriebspanels beteiligen, vertrauen darauf, dass ihre Angaben vertraulich behandelt werden. Die Gewährleistung der Anonymität der befragten Betriebe ist eine wichtige Voraussetzung, um mögliche Kandidaten für eine Teilnahme motivieren zu können. Existieren Zweifel, ob der zugesicherte Datenschutz eingehalten werden kann, besteht die Gefahr, dass die Befragten entweder überhaupt nicht mehr bereit sind, an der Befragung teilzunehmen, oder absichtlich falsche Angaben machen, um ihre Identität zu verschleiern. Die Konsequenzen für die Qualität und den Nutzen der erhobenen Daten können in diesem Fall katastrophal sein.

Andererseits besteht ein starkes wissenschaftliches Interesse, die erhobenen Daten einer möglichst breiten Fachöffentlichkeit zur Verfügung zu stellen und damit die wissenschaftliche Forschung anzuregen. Daher bietet das FDZ seit einigen Jahren externen Wissenschaftlern die Möglichkeit, per Datenfernverarbeitung oder im Rahmen eines Gastaufenthalts, mit den Datensätzen des Betriebspanels zu rechnen. Bevor die Auswertungen zum Beispiel für Publikationen verwendet werden können, muss eine Datenschutzprüfung durch die Mitarbeiter des FDZ durchgeführt werden. Beide Verfahren sind allerdings sowohl für die externen Wissenschaftler als auch für das FDZ mit einem nicht unerheblichen Aufwand und hohen Kosten verbunden. Bei der Datenfernverarbeitung bereiten die Wissenschaftler ihre Computeranalysen anhand eines Testdatensatzes vor und schicken den fertig gestellten Programmcode an die Mitarbeiter des Forschungsdatenzentrums, die die Analysen über den Originaldatensatz laufen lassen und die Ergebnisse an die Wissenschaftler zurückschicken. Dieser Testdatensatz bietet zwar den gleichen Variablenumfang, allerdings werden die Originalwerte verändert, um den Datenschutz zu gewährleisten. Eine Beschreibung der Testdatensätze für das IAB-Betriebspanel findet sich in Jacobebbinghaus/Müller/Orban (2010).

Beim Fernrechnen kann es passieren, dass die Analysecodes mehrfach zwischen den Forschern und den Mitarbeitern des Forschungsdatenzentrums hin- bzw. hergeschickt und revidiert werden müssen, bevor das gewünschte Ergebnis vorliegt. Alle Zwischenergebnisse müssen von Mitarbeitern des Forschungsdatenzentrums aufwendig überprüft werden, um sicherzustellen, dass bei Bereitstellung der Ergebnisse der Datenschutz nicht verletzt wird.

Der Gastaufenthalt bietet Wissenschaftlern die Möglichkeit, ihre Analysen direkt mit den Originaldaten durchzuführen. Allerdings ist dieses Vorgehen insbesondere für ausländische Gäste mit erheblichen Reisekosten und Aufwand verbunden, da es oft etliche Tage dauert, bis die Datensätze soweit aufbereitet sind, dass die gewünschten Analysen durchgeführt werden können. Außerdem besteht lediglich ein eingeschränkter Datenzugang und auch bei diesem Vorgehen müssen Auswertungen nachträglich von den Mitarbeitern des Forschungsdatenzentrums auf Einhaltung des Datenschutzes geprüft werden. Um diesen Zeit- und Kostenaufwand für beide Seiten zu reduzieren, werden in der Literatur zahlreiche

Verfahren diskutiert, die eine Herausgabe der Daten in anonymisierter Form ermöglichen. Neben der Möglichkeit des data swapping, bei dem die Informationen in einzelnen Zellen untereinander ausgetauscht werden, findet der Ansatz der Mikro-Aggregation häufige Anwendung. Bei diesem Ansatz werden mehrere Zellen mit hohem Identifikationsrisiko (z. B. die Beschäftigtenzahl bei sehr großen Unternehmen) zusammengefasst werden und die einzelnen Werte z. B. durch Mittelwerte ersetzt werden. Weitere Alternativen wären beispielsweise das Weglassen einzelner sensibler Variablen oder das Hinzufügen von Störtermen zu jeder Beobachtung einzelner Variablen (Noise Addition).

Alle Ansätze sind aber mit einem mehr oder weniger großen Verlust an Information verbunden. Außerdem ist in vielen Fällen für eine spätere Analyse die genaue Kenntnis der verwendeten Anonymisierungstechnik und/oder die Verwendung von Spezialsoftware Voraussetzung, um zu validen Analyseergebnissen zu kommen. Ein sehr innovativer und in Europa noch relativ unbekannter Ansatz zur Anonymisierung ist die Erzeugung synthetischer Datensätze, die erstmals 1993 von Rubin vorgeschlagen wurde. Bei diesen Verfahren werden einzelne Beobachtungen (teilweise synthetische Daten) oder sogar der komplette Datensatz (vollständig synthetische Daten) durch künstlich erzeugte Werte ersetzt. Der entscheidende Vorteil bei diesem Verfahren liegt in der Universalität des Ansatzes. Filterstrukturen und Restriktionen im Datensatz wie Nicht-Negativität können bei der Erstellung berücksichtigt werden. Außerdem ist der Ansatz auf kontinuierliche Variablen ebenso anwendbar wie auf kategoriale Variablen. Im Gegensatz zu vielen anderen Verfahren wird zudem versucht, die gemeinsame Verteilung der Variablen des Datensatzes zu erhalten und nicht nur einzelne Statistiken.

2 Was sind synthetische Daten?

Das Verfahren zur Erzeugung synthetischer Daten ist eng mit dem Verfahren der multiplen Imputation (Rubin, 1987) verwandt. Ähnlich wie bei der multiplen Imputation werden für jede Variable Regressionsmodelle basierend auf den Originaldaten geschätzt. Die geschätzten Parameter werden aber nicht dazu verwendet, fehlende Werte zu ergänzen, sondern um die ursprünglich beobachteten Werte durch imputierte (=synthetische) Werte zu ersetzen. Analog zur multiplen Imputation müssen mehrere synthetische Datensätze erzeugt werden, um eine valide Varianzschätzung zu ermöglichen, da die zusätzliche Unsicherheit, die sich aus der Tatsache ergibt, dass es sich bei den imputierten Werten lediglich um geschätzte Werte handelt, berücksichtigt werden muss.

Je nachdem, ob es sich um partiell synthetische Daten oder vollständig synthetische Daten, bei denen in einem ersten Schritt die fehlenden Werte imputiert wurden, handelt, ergeben sich unterschiedliche Kombinationsregeln, um aus den einzelnen Datensätzen das endgültige Analyseergebnis zu erzielen. Diese Kombinationsregeln unterscheiden sich von den Kombinationsregeln bei der Ergänzung fehlender Werte durch multiple Imputation. Einen guten Überblick über die verschiedenen Einsatzmöglichkeiten der multiplen Imputation und die jeweils resultierenden Kombinationsregeln geben Reiter/Raghunathan (2007).

Prinzipiell lassen sich voll synthetische und partiell synthetische Datensätze unterschei-

den. Zur Erstellung vollständig synthetischer Datensätze sind in jedem Fall zusätzliche Variablen nötig, die für die Grundgesamtheit ausnahmslos beobachtet vorliegen müssen. Hier können beispielsweise Registerdaten verwendet werden. Um nun synthetische Datensätze zu erzeugen, werden neue Stichproben aus dieser Grundgesamtheit gezogen. Für diese Stichproben werden die Variablen aus der Befragung als fehlende Werte betrachtet und mit dem Ansatz der multiplen Imputation ergänzt. Die so gewonnenen imputierten Datensätze können dann als synthetische Datensätze der Öffentlichkeit zugänglich gemacht werden.

Da es sich bei den imputierten Werten um künstliche Werte für Betriebe, die nie an der Befragung teilgenommen haben, handelt, ist eine Identifizierung einzelner Betriebe anhand der ergänzten Daten nahezu ausgeschlossen. Ein weiterer bedeutender Vorteil dieses Ansatzes liegt in der leichteren Analyse der ergänzten Datensätze. Diese können als eine einfache Zufallsstichprobe aus der Grundgesamtheit erzeugt werden, während viele Umfragen ein kompliziertes Stichprobendesign verwenden, das bei den späteren Analysen berücksichtigt werden muss. Außerdem ist das zur Verfügung stehende Analysewerkzeug für reine Zufallsstichproben wesentlich umfangreicher.

Im Kontrast dazu werden bei partiell synthetischen Datensätzen (Little, 1993) nur diejenigen Variablen ersetzt, die zu Identifizierungszwecken verwendet werden können (Schlüsselvariablen) oder die besonders sensible Informationen enthalten. Schlüsselvariablen sind dabei diejenigen Variablen, die auch in allgemein zugänglichen Quellen verfügbar sind und somit ein hohes Identifikationsrisiko bergen. Zum Beispiel kann die Angabe der Betriebsgröße leicht dazu führen, dass gerade große Unternehmen anhand dieser Variablen eindeutig identifiziert werden können und somit auch sensiblere Daten eindeutig diesem Unternehmen zugeordnet werden können. Alle übrigen Variablen werden nicht verändert. Entsprechend werden für diesen Ansatz auch keine neuen Stichproben benötigt. Vielmehr umfasst der partiell synthetische Datensatz genau die Einheiten, die an der Befragung teilgenommen haben.

Ein deutlicher Vorteil vollständig synthetischer Datensätze liegt darin, dass absolut keine tatsächlich beobachtete Information veröffentlicht wird. Einerseits beinhaltet der veröffentlichte Datensatz ausschließlich künstliche Werte, andererseits werden diese Werte für Einheiten erzeugt, die gar nicht an der Befragung teilgenommen haben. Ein potenzieller Datenangreifer kann also sein Wissen darüber, ob ein Individuum oder ein Betrieb an einer Befragung teilgenommen hat, nicht nutzen. Aus der Perspektive der Datensicherheit bietet dieser Ansatz somit den größtmöglichen Schutz. Andererseits sind aber auch einige schwerwiegende Nachteile mit diesem Ansatz verbunden: Zum einen müssen vollständig beobachtete Daten für die Grundgesamtheit verfügbar sein. Zum anderen muss eine eindeutige Zuordnung der Befragungsteilnehmer zu den Informationen über die Grundgesamtheit möglich sein. Außerdem ist die Erzeugung vollständig synthetischer Daten mit einem sehr großen Aufwand verbunden. Das abwechselnde Ziehen neuer Stichproben und anschließende Imputieren kann sehr zeit- und rechenintensiv werden.

Der größte Nachteil ist aber in der hohen Abhängigkeit von der Qualität der gewählten Imputationsmodelle zu sehen. In vielen Fällen ist die Entwicklung brauchbarer Modelle sehr

aufwendig und für einzelne Variablen kann sich die Erstellung eines Modells, das zuverlässige Vorhersagen ermöglicht, als nahezu unmöglich erweisen. Wenn aber von den betroffenen Variablen kein Identifizierungsrisiko ausgeht, bzw. die Variablen keine sensiblen Informationen enthalten, stellt sich die Frage, warum diese Variablen überhaupt imputiert werden müssen. Zumal eine ”schlecht” imputierte Variable auch negative Konsequenzen auf andere Variablen hat, da diese Variable als Prädiktor bei der Erzeugung andere Variablen verwendet wird.

Genau hierin liegt der deutliche Vorteil teilweise synthetischer Datensätze. Es werden nur diejenigen Variablen imputiert, die ein Risiko darstellen. Das kann die Abhängigkeit von der Qualität des gewählten Modells deutlich senken, vor allem wenn man bedenkt, dass eine Vielzahl häufig binärer Variablen zwar von hohem wissenschaftlichen Interesse sein können, selbst aber keinerlei oder ein vernachlässigbar kleines Identifikationsrisiko darstellen. Diese Variablen können aber andererseits eine sehr hohe Qualität als Prädiktoren entfalten, so dass sich auch kleine Fehlspezifikationen negativ auf die Imputation anderer Variablen auswirken können. Insofern ist es wünschenswert, diese Variablen unverändert im Datensatz zu belassen. Da weniger Variablen zu imputieren sind, reduziert sich entsprechend die Anzahl der sehr arbeitsaufwendigen Modellspezifikationen. Außerdem entfällt der Zwischenschritt der Generierung neuer Stichproben.

Auf der anderen Seite ist das verbleibende Identifizierungsrisiko bei teilweise synthetischen Datensätzen sicher höher als bei vollständig synthetischen Datensätzen. Zum einen verbleiben Variablen unverändert im Datensatz, andererseits beinhaltet der veröffentlichte Datensatz ausschließlich die ursprünglichen Befragungsteilnehmer. Aus diesen Gründen ist eine sehr genaue Analyse des verbleibenden Identifizierungsrisikos für teilweise synthetische Daten unabdingbar. Es muss sichergestellt sein, dass keine Variablen unverändert veröffentlicht werden, die ein Risiko darstellen, und dass die Veränderungen an den risikobehafteten Variablen stark genug sind, wodurch eine zuverlässige Reidentifikation ausgeschlossen ist.

Aufgrund seiner hohen Flexibilität und Anwendbarkeit, auch für sehr komplexe, zusammengespielte Paneldatensätze, wird der Ansatz zur Erzeugung teilweise synthetischer Daten in den letzten Jahren international immer stärker eingesetzt. Seit mehreren Jahren arbeitet eine Forschergruppe um Prof. John Abowd (Cornell University) im Auftrag des U.S. Census Bureaus daran, ein Public Use File des Survey of Income and Programm Participation auf Grundlage synthetischer Daten zu erzeugen. Umfangreiche Untersuchungen haben die hohe Datenqualität und Datensicherheit bestätigt (Abowd/Stinson/Benedetto, 2006). Eine erste Version der synthetisierten Daten wurde der Öffentlichkeit 2007 zugänglich gemacht. Das Projekt On the Map des U.S. Census Bureaus stellt die derzeit erfolgreichste Anwendung synthetischer Daten dar. Auf den Internetseiten des U.S. Census Bureaus kann sich jeder Nutzer Berufspendlerströme in den gesamten USA sehr detailliert grafisch anzeigen lassen. Die zugrunde liegenden Daten wurden durch Erzeugung synthetischer Datensätze anonymisiert (Machanavajjhala u. a., 2008). Weitere synthetische Datensätze sind derzeit in den USA in Entwicklung (The Longitudinal Business Database, The Longitudinal Employer-Household Dynamics Survey, The American Communities Survey).

Neben Forschern in den USA befassen sich auch Wissenschaftler in Kanada, Neuseeland (Graham/Penny, 2005; Graham/Young/Penny, 2009) und Australien mit der Erzeugung synthetischer Datensätze. Das vom Bundesministerium für Bildung und Forschung geförderte Drittmittelprojekt FAWE Panel hat unter anderem gezeigt, dass dieser Ansatz auch für das IAB Betriebspanel zu sehr guten Ergebnissen führen kann (Drechsler u. a., 2008; Drechsler/Bender/Rässler, 2008). Mit der Bereitstellung des IAB Betriebspanels wird der europaweit erste synthetische Datensatz interessierten Forschern zur Verfügung gestellt. Die europäische Union hat unlängst ein Gutachten über die Anwendbarkeit der synthetischen Daten für die Statistiken der EU in Auftrag gegeben (Domingo-Ferrer/Drechsler/Polettini, 2009).

3 Hinweise zur Nutzung synthetischer Datensätze

Generell werden in den synthetischen Datensätzen nur die statistischen Zusammenhänge korrekt wiedergegeben, die bei der Erstellung des Imputationsmodells berücksichtigt wurden. Für Variablen, die nicht im der Imputation zugrunde liegenden Regressionsmodell berücksichtigt wurden, wird implizit eine bedingte Unabhängigkeit angenommen. Das heißt, implizit wird unterstellt, dass gegeben allen anderen Variablen im Regressionsmodel kein Zusammenhang zwischen diesen Variablen und der zu erklärenden Variablen besteht.

Verwendet nun der Datennutzer ein Analysemodell, das erklärende Variablen enthält, die nicht im Imputationsmodell berücksichtigt wurden, kann es zu Verzerrungen kommen, wenn die oben genannte bedingte Unabhängigkeitsannahme nicht erfüllt ist. Daher werden bei einem Imputationsmodell im Gegensatz zu einem ökonomischen Modell nach Möglichkeit immer alle Variablen des Datensatzes als erklärende Variablen aufgenommen. Allerdings ist es aus verschiedenen Gründen nicht immer möglich, alle Variablen zu berücksichtigen. Im Betriebspanel werden beispielsweise bestimmte Fragen nur von einem Teil der Befragten beantwortet. Sollten all diese Variablen mit berücksichtigt werden, ließen sich die Modelle nur mit einem sehr kleinen Teil des gesamten Datensatzes schätzen (nur für die Betriebe, die tatsächlich alle Fragen im Fragebogen beantwort haben), was im besten Fall nur zu einem Verlust an Effizienz führt, im schlechtesten aber auch wahrscheinlicheren Fall hingegen verzerrte Ergebnissen zur Folge hat, wenn sich die nicht berücksichtigten Betriebe systematisch von den Betrieben unterscheiden, die dem Modell zugrunde liegen. Daher wurden bestimmte Variablen, die nur ein Teil der Befragten beantworteten, nicht berücksichtigt. Allerdings mussten längst nicht alle derartige Variablen aus dem Modell genommen werden, da es bei einem großen Teil der Fragen offensichtlich ist, dass eine Nichtbeantwortung mit einer Null gleichzusetzen ist. So gibt es im Betriebspanel beispielsweise die Frage, ob der Betrieb am Stichtag Teilzeitbeschäftigte hatte. Falls ein Betrieb diese Frage mit ”nein” beantwortet, wird im die nächste Frage nach der Anzahl der Teilzeitbeschäftigten nicht gestellt. Es ist jedoch offensichtlich, dass der Betrieb keine Teilzeitbeschäftigte hat und somit eine Null eingesetzt werden kann. Derartige Filterfragen sind im Betriebspanel zahlreich vertreten.

Andererseits können Multikollinearitätsprobleme und geringe Fallzahlen dazu führen, dass für einzelne Variablen weitere erklärende Variablen aus dem Modell genommen werden

mussten. Um dem Nutzer eine Einschätzung zu ermöglichen, ob sich für die geplante Analyse mit den synthetischen Daten valide Ergebnisse erzielen lassen, ist in der Datei "Modellbeschreibungen" aufgelistet, welche erklärenden Variablen bei den einzelnen anonymisierten Variablen aus dem Modell genommen werden mussten. Allerdings dürften die negativen Auswirkungen der Reduktion des Imputationsmodells bei Multikollinearität eher gering ausfallen, denn Multikollinarität bedeutet letztendlich, dass ein starker linearer Zusammenhang zwischen der aus dem Modell genommenen Variable und einem anderen Prädiktor, der im Modell verbleibt, besteht. Somit scheint die Annahme der bedingten Unabhängigkeit für die aus dem Modell genommene Variable gerechtfertigt.

Vorsicht ist aber geboten, wenn eine zu analysierende Variable aus mehreren bestehenden Variablen gebildet wird und kleine Veränderungen in den zugrunde liegenden Variablen sehr starke Auswirkungen auf die gebildete Variable haben (siehe dazu auch das unten beschriebene Beispiel). Da die Anonymisierung zwangsläufig zur Folge hat, dass individuelle Beobachtungen verändert werden, kann es dann zu Verzerrungen bei der Analyse kommen.

4 Analysen mit synthetischen Datensätzen[1]

Viele potenzielle Datennutzer werden sich an dieser Stelle fragen, wie eine korrekte Analyse mit synthetischen Datensätzen durchzuführen ist. In diesem Kapitel werden einige Hinweise zum Umgang mit synthetischen Datensätzen und zur richtigen Verwendung der Kombinationsregeln gegeben mit denen die Ergebnisse über die 25 Datensätze gepoolt werden.

Wir empfehlen, dass der Nutzer zuerst mit einem einzelnen Datensatz beginnt und für diesen Datensatz seinen Datenaufbereitungs- und Analysecode schreibt. Für diese Zwecke steht der Datensatz *iabbp_2007_single* zur Verfügung. Da alle synthetischen Datensätze genau gleich aufgebaut sind, lässt sich der entwickelte Code dann problemlos auf die anderen Datensätze übertragen (alle Datensätze sind in dem File *iabbp_2007_all* gespeichert). Demnach unterscheiden sich die synthetischen Datensätze in der Analyse nicht von jedem anderen Mikrodatensatz. Der einzige Unterschied besteht darin, dass die Analysen mehrmals durchgeführt werden. Wir empfehlen die Ergebnisse der Analysen jeweils in eigenen Datensätzen zu speichern. Das erleichtert die Kombination der einzelnen Auswertungen, um das endgültige Ergebnis zu erhalten. Wir empfehlen außerdem, dass alle Entscheidungen bzgl. der statistischen Inferenz nur unter Verwendung aller Datensätze und der korrekten Kombinationsregeln (s.u.) getroffen werden. D.h. wir raten davon ab, dass ein einzelner Datensatz dazu verwendet wird, über die Modellspezifikation zu entscheiden. Die Theorie, die der Erzeugung synthetischer Datensätze zugrunde liegt, impliziert, dass mehrere Analysen mit mehreren unabhängig erzeugten Datensätzen notwendig sind, um die Unsicherheit in den Originaldaten korrekt abzubilden.

Bei der Erstellung der synthetischen Datensätze für das IAB Betriebspanel wurde ein zweistufiges Verfahren verwendet. Zunächst wurden alle fehlenden Werte im Originaldatensatz

[1] Dieses Kapitel ist in weiten Abschnitten aus Abowd/Stinson/Benedetto (2006) übernommen.

mehrfach imputiert. Danach wurden für jeden dieser imputierten Datensätze mehrere synthetische Datensätze erzeugt. Dieses zweistufige Verfahren muss auch bei der Analyse berücksichtigt werden. Jeder Datensatz enthält zwei Variablen, die die Beziehung zwischen den Datensätzen widerspiegeln. Die Variable *_m* zeigt an, welcher vollständig imputierte Datensatz als Ausgangspunkt für die Erzeugung des synthetischen Datensatzes gedient hat. Die Variable *_r* gibt die Nummer des synthetischen Datensatzes an. Insgesamt gibt es 5 vollständig imputierte Datensätze, entsprechend nimmt *_m* die Werte 1 bis 5 an. Für jeden vollständig imputierten Datensatz wurden jeweils 5 synthetische Datensätze erzeugt. Entsprechend nimmt die Variable *_r* ebenfalls die Werte 1 bis 5 an. Anhand dieser Information kann der Nutzer identifizieren, welcher synthetische Datensatz von welchem vollständig imputierten Datensatz stammt. Diese Information ist für die Verwendung der Kombinationsregeln notwendig.

Jede Auswertung kann mit den synthetischen Daten durchgeführt werden, indem sie mit jedem synthetischen Datensatz einzeln durchgeführt wird. Das Endergebnis ergibt sich dann als Mittelwert der Ergebnisse der einzelnen Datensätze. Ist man beispielsweise am durchschnittlichen Umsatz aller Unternehmen einer Branche interessiert, berechnet man diesen Umsatz für jeden einzelnen Datensatz mit Standardmethoden und mittelt am Ende die Ergebnisse über die 25 synthetischen Datensätze. Auf diese Weise lassen sich alle Arten von Punktschätzern egal ob Mittelwerte, Regressionskoeffizienten oder Totalwerte berechnen.

Die Berechnung der geschätzten Gesamtvarianz, die zum Beispiel zur Berechnung von Konfidenzintervallen oder Teststatistiken verwendet werden kann, ist etwas komplizierter, kann aber nach wie vor mit jeder Standardsoftware bestimmt werden. Zur Berechnung der Gesamtvarianz sind folgende Formeln notwendig:

$$\bar{b}_m = \sum_{l=1}^{m}\sum_{i=1}^{r}(q_i^{(l)} - \bar{q}^{(l)})^2/m(r-1) = \sum_{l=1}^{m} b^{(l)}/m \quad (1)$$

$$B_M = \sum_{l=1}^{m}(\bar{q}^{(l)} - \bar{q}_M)^2/(m-1) \quad (2)$$

$$\bar{u}_M = \sum_{i=1}^{m}\sum_{i=1}^{r} u_i^{(l)}/(mr) \ , \quad (3)$$

wobei $q_i^{(l)}$ und $u_i^{(l)}$ den Punktschätzer und den Schätzer für dessen Varianz im i-ten synthetischen Datensatz des l-ten imputierten Datensates darstellt, $\bar{q}^{(l)}$ den Durchschnitt der 5 synthetischen Datensätze, die aus dem l-ten imputierten Datensatz generiert wurden, bildet und $\bar{q}_M = \sum_{i=1}^{r} q_i^{(l)}/(mr)$ dem Durchschnitt über alle $q_i^{(l)}$ entspricht.

Die Gesamtvarianz des Punktschätzers berechnet sich dann als:

$$T_M = (1 + 1/m)B_M - \bar{b}_M/r + \bar{u}_M. \quad (4)$$

D.h. neben dem Punktschätzer $q_i^{(l)}$ der interessierenden Teststatistik sollte immer auch

die zugehörige geschätzte Varianz $u_i^{(l)}$ des Punktschätzers (bei Regressionskoeffizienten also beispielsweise das Quadrat des Standardfehlers) für jeden synthetischen Datensatz gespeichert werden.[2] Um die Gesamtvarianz zu ermitteln, muss zunächst die durchschnittliche Varianz $\bar{u}_M$ innerhalb der Datensätze ermittelt werden, in dem über alle $u_i^{(l)}$ gemittelt wird. Danach müssen die Indikatorvariablen *_m* und *_r* verwendet werden, um die Varianz zwischen den imputierten Datensätzen B_M und die durchschnittliche Varianz zwischen den synthetischen Datensätzen $\bar{b}_m$ nach den Formeln (2) und (3) zu berechnen. Im ersten Schritt wird die Varianz $b^{(l)} = \sum_{i=1}^{r}(q_i^{(l)} - \bar{q}^{(l)})^2$ zwischen den 5 synthetischen Datensätzen, denen der gleiche imputierte Datensatz (*_m* ist also fix) zugrunde liegt, berechnet. Da es insgesamt 5 (*_m=1,...5*) imputierte Datensätze gibt, muss diese Varianz 5 mal bestimmt werden. Anschließend wird die durchschnittliche Varianz ermittelt, indem über die 5 berechneten Varianzen gemittelt wird: $\bar{b}_m = \sum_{l=1}^{m} b^{(l)}/m$. Im letzten Schritt muss noch die Varianz zwischen den imputierten Datensätzen berechnet werden, indem die Varianz von $\bar{q}^{(l)}$ um den Gesamtmittelwert $\bar{q}_M$ bestimmt wird. Wurden die Ergebnisse der Analysen der einzelnen synthetischen Datensätze jeweils in eigenen Datensätzen gespeichert, lassen sich die einzelnen Komponenten zur Bestimmung der Gesamtvarianz einfach ermitteln.

Um Konfidenzintervalle zu bestimmen, kann vereinfachend eine Normalverteilung angenommen werden und beispielsweise das 95%-Konfidenzintervall als $\bar{q}_M \pm 1.96\sqrt{T_m}$ ermittelt werden. Für eine exakte Bestimmung sollte jedoch eine t-Verteilung statt der Normalverteilung verwendet werden, wobei sich die Freiheitsgrade nach folgender Formel berechnen:

$$\nu_M = \left(\frac{((1+1/m)B_M)^2}{(m-1)T_M^2} + \frac{(\bar{b}_m/r)^2}{m(r-1)T_M^2} \right)^{-1} \tag{5}$$

In seltenen Fällen kann es passieren, dass T_M negativ wird. In diesem Fall emfehlen wir den konservativen Varianzschätzer $T_M^{adj} = (1+1/m)B_m + \bar{u}_M$ und für die Berechnung der Freiheitsgrade der t-Verteilung:

$$\nu_M^{adj} = (m-1)(1 + m\bar{u}_M/((m+1)B_M)). \tag{6}$$

5 Ergebnisse von Beispielanalysen

Im folgenden Kapitel sollen zwei Beispielanalysen einerseits die hohe Datenqualität der synthetischen Daten veranschaulichen, andererseits aber auch beispielhaft aufführen, bei welchen Analysen Vorsicht geboten ist. Um einen fairen Vergleich zu gewährleisten, werden die Ergebnisse der synthetischen Datensätze, bei dem alle fehlenden Werte mehrfach imputiert wurden, stets mit den Ergebnissen des Betriebspanels verglichen. Um die Ergebnisse besser vergleichen zu können, wird neben der Gegenüberstellung der Punktschätzer auch ausgewiesen, wie stark sich die 95%-Konfidenzintervalle der Punktschätzer überlap-

[2] Die Berechnungen müssen in jedem Fall mit den Varianzen und nicht mit den Standardabweichungen durchgeführt werden. Um die endgültige Standardabweichung oder den Standardfehler zu berechnen, muss die Wurzel aus der berechneten Gesamtvarianz gezogen werden.

pen (vgl. Karr et al., 2006). Das Maß liegt immer zwischen 0 und 1, wobei 0 keine Überlappung, eine 1 eine exakte Überlappung der Konfidenzintervalle widerspiegelt. Dieses Maß ermöglicht einen aussagekräftigeren Vergleich, da Punktschätzer mit einem großen Standardfehler durchaus sehr weit auseinander liegen können, bei hoher Überlappung aber trotzdem auf Basis der synthetischen Datensätze ähnliche Rückschlüsse möglich sind wie mit den Originaldaten. Umgekehrt können bei sehr kleinen Standardfehlern nahe beieinander liegende Punktschätzer in ihrer statistischen Inferenz trotzdem eine schlechte Qualität aufweisen.

Bei der ersten Regression gibt die abhängige Variable wieder, ob ein Betrieb Teilzeitbeschäftigte hat oder nicht. Die 19 erklärenden Variablen umfassen unter anderem Betriebsgrößendummies, ob Veränderungen in der Beschäftigtenzahl erwartet werden und verschiedene Informationen zur Personalstruktur. Die Regression wird für West- und Ostdeutschland unabhängig durchgeführt.

Tabelle 1: Regressionsergebnisse einer Probit Regression von *Teilzeitbeschäftigten(ja/nein)* auf 19 erklärende Variablen in Westdeutschland.

	original data	synth. data	CI overlap	z-score org.	z-score syn
Achsenabschnitt	-0.809	-0.752	0.87	-7.23	-6.85
5-10 Beschäftigte	0.443	0.437	0.97	8.52	7.99
10-20 Beschäftigte	0.658	0.636	0.90	11.03	10.88
20-50 Beschäftigte	0.797	0.785	0.95	13.02	12.36
100-200 Beschäftigte	0.892	0.908	0.96	9.23	9.48
200-500 Beschäftigte	1.131	1.125	0.99	9.99	9.87
>500 Beschäftigte	1.668	1.641	0.97	8.22	8.33
Besch.wachstum erwartet	0.010	0.006	0.98	0.18	0.12
Besch.rückgang erwartet	0.087	0.100	0.96	1.11	1.27
Anteil Frauen	1.449	1.366	0.73	17.63	18.71
Anteil Hochqualifizierte	0.319	0.368	0.91	2.18	2.59
Anteil Geringqualifizierte	1.123	1.148	0.93	12.17	11.87
Anteil befristet Beschäftigte	-0.327	-0.138	0.75	-1.74	-0.71
Anteil Leiharbeiter	-0.746	-0.856	0.88	-3.09	-4.24
Einstellungen (6 Monate)	0.394	0.369	0.87	8.33	7.82
Entlassungen (6 Monate)	0.294	0.279	0.92	6.38	6.03
ausländisches Eigentum	-0.113	-0.117	0.99	-1.33	-1.38
gute/sehr gute Ertragslage	0.029	0.033	0.98	0.72	0.82
Zahlung über Tarifvertrag	0.020	0.031	0.95	0.35	0.54
Branchentarifvertrag	0.016	0.007	0.95	0.31	0.13

Tabelle 2: Regressionsergebnisse einer Probit Regression von *Teilzeitbeschäftigten(ja/nein)* auf 19 erklärende Variablen in Ostdeutschland.

	original data	synth. data	CI over-lap	z-score org.	z-score syn
Achsenabschnitt	-0.712	-0.742	0.93	-6.42	-7.21
5-10 Beschäftigte	0.266	0.257	0.96	4.81	4.53
10-20 Beschäftigte	0.416	0.399	0.93	6.94	6.76
20-50 Beschäftigte	0.542	0.532	0.96	9.18	8.72
100-200 Beschäftigte	0.757	0.808	0.86	8.02	8.47
200-500 Beschäftigte	0.971	1.013	0.91	8.25	8.57
>500 Beschäftigte	1.401	1.422	0.98	5.69	5.66
Besch.wachstum erwartet	-0.041	-0.040	1.00	-0.73	-0.73
Besch.rückgang erwartet	0.035	0.040	0.98	0.44	0.50
Anteil Frauen	1.006	1.041	0.88	12.63	14.93
Anteil Hochqualifizierte	0.221	0.197	0.95	1.86	1.76
Anteil Geringqualifizierte	0.976	1.042	0.87	8.44	7.84
Anteil befristet Beschäftigte	-0.049	0.049	0.84	-0.31	0.34
Anteil Leiharbeiter	-0.176	-0.232	0.94	-0.73	-1.08
Einstellungen (6 Monate)	0.230	0.210	0.89	4.95	4.55
Entlassungen (6 Monate)	0.301	0.295	0.97	6.43	6.35
ausländisches Eigentum	-0.176	-0.176	1.00	-1.83	-1.84
gute/sehr gute Ertragslage	0.097	0.097	1.00	2.35	2.37
Zahlung über Tarifvertrag	0.080	0.086	0.98	1.04	1.10
Branchentarifvertrag	0.097	0.069	0.86	1.87	1.36

Die Ergebnisse zeigen deutlich die hohe Datenqualität. Alle Punktschätzer liegen sehr nah an den Punktschätzern der Originaldaten und die Überlappung des Konfidenzintervall liegt bei mehr als 90% für die meisten Koeffizienten. Auch die t-Werte liegen sehr nah an denen der Originaldaten, so dass die Analyse mit den synthetischen Daten die gleichen Rückschlüsse zulässt wie die Analyse der Originaldaten.

Im Folgenden soll ein Beispiel die Grenzen der synthetischen Daten aufzeigen. Vor der geplanten Analyse sollte sich jeder Wissenschaftler überlegen, ob die synthetischen Daten für die gewählte Analyse plausible Ergebnisse erwarten lassen können. Als ein fiktives Beispiel für eine Analyse, die keine plausiblen Ergebnisse erzielt, wird in der im vorangegangenen Abschnitt beschriebenen Analyse die abhängige Variable durch eine Variable ersetzt, die die Beschäftigungsentwicklung widerspiegelt. Die Variable nimmt den Wert eins an, wenn die Beschäftigtenzahl zwischen 2006 und 2007 gestiegen ist. Tabelle 3 gibt vergleichbare Ergebnisse für Westdeutschland wieder.

Es lässt sich leicht erkennen, dass die Ergebnisse der synthetischen Datensätze hier sehr stark von den Ergebnissen der Originaldatensätze abweichen. Die Ursache hierfür ist in der Modellierung der abhängigen Variable zu suchen. In Grafik 1 sind jeweils QQ-Plots der synthetischen gegen die Originaldaten dargestellt. Man erkennt deutlich, dass die synthetischen Daten die Verteilung der Originaldaten sehr gut nachbilden. Das gilt sowohl für die Beschäftigtenzahlen in den Jahren 2006 und 2007 als auch für die Differenz in der Beschäftigtenzahl zwischen 2006 und 2007. Der große Unterschied in den Ergebnissen resultiert also nicht aus Imputationsmodellen, die die wahre Verteilung schlecht widerspiegeln. Das Problem liegt viel mehr darin, dass sich die Beschäftigtenzahlen zwischen den beiden Jahren kaum verändert haben. Mehr als 5.000 Betriebe berichten im Originaldatensatz überhaupt keine Veränderung in der Beschäftigtenentwicklung und mehr als 90% der Betriebe geben an, dass die Veränderung bei $\pm 5\%$ lag. Daher ist es leicht möglich, dass beispielsweise ein Betrieb im Originaldatensatz in einem Jahr 50 Beschäftigte und im nächsten Jahr 52 Beschäftigte angegeben hat. Im synthetischen Datensatz hat dieser Betrieb dann beispielsweise 53 Beschäftigte im ersten Jahr und 52 Beschäftigte im zweiten Jahr. Somit liegen die imputierten Werte sehr nahe an den Originalwerten, aber die leichten Veränderungen haben dazu geführt, dass die Variable Beschäftigungsentwicklung von 1 auf 0 wechselt.

Tabelle 3: Regressionsergebnisse einer Probit Regression des *Beschäftigungstrends* auf 19 erklärende Variablen in Westdeutschland.

	original data	synth. data	CI over-lap	z-score org.	z-score syn	CI length ratio
Achsenabschnitt	-1.396	-0.978	0.05	-11.99	-9.28	0.92
5-10 Beschäftigte	0.130	0.354	0.00	2.61	7.75	0.92
10-20 Beschäftigte	0.316	0.495	0.05	6.19	11.19	0.87
20-50 Beschäftigte	0.355	0.541	0.05	7.33	10.93	1.06
100-200 Beschäftigte	0.366	0.351	0.94	5.69	6.09	0.91
200-500 Beschäftigte	0.475	0.347	0.48	7.29	5.80	0.92
>500 Beschäftigte	0.375	0.472	0.66	5.06	6.58	0.99
Besch.wachstum erwartet	0.374	0.148	0.00	9.29	3.59	1.05
Besch.rückgang erwartet	-0.376	-0.020	0.00	-6.16	-0.38	0.86
Anteil Frauen	-0.140	-0.054	0.67	-2.09	-0.84	1.00
Anteil Hochqualifizierte	0.229	0.199	0.91	1.94	2.05	0.83
Anteil Geringqualifizierte	-0.043	-0.004	0.84	-0.68	-0.07	0.97
Anteil befristet Beschäftigte	0.434	0.226	0.62	3.25	1.60	1.07
Anteil Leiharbeiter	0.058	0.013	0.69	0.94	0.08	2.61
Einstellungen (6 Monate)	0.948	0.368	0.00	24.94	11.60	0.84
Entlassungen (6 Monate)	-0.172	-0.030	0.00	-4.42	-0.97	0.81
ausländisches Eigentum	-0.165	-0.113	0.79	-2.60	-1.90	0.98
gute/sehr gute Ertragslage	0.248	0.100	0.00	7.69	3.35	0.93
Zahlung über Tarifvertrag	0.039	0.033	0.96	0.87	0.81	0.91
Branchentarifvertrag	0.003	0.063	0.62	0.06	1.72	0.85

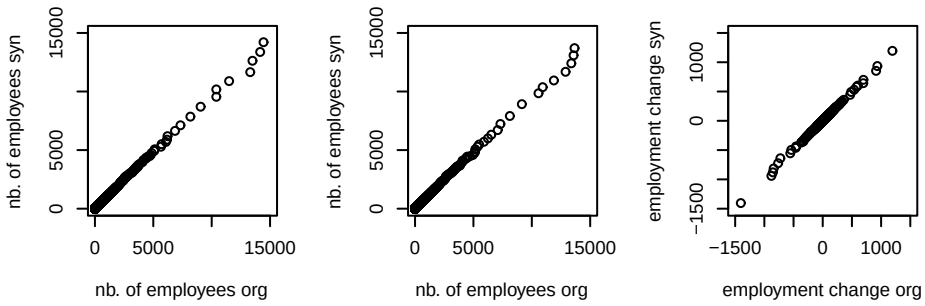

Abbildung 1: QQ-plots der Anzahl der Beschäftigten 2006 und 2007 und der Beschäftigungsentwicklung zwischen den zwei Jahren.

Somit kann also eine leichte Veränderung der Daten eine sehr große Veränderung der neu gebildeten Variable Beschäftigungsentwicklung haben. In diesem Zusammenhang ist es also sehr wichtig, dass sich jeder Nutzer bei der Erzeugung eigener Variablen aus dem vorhandenen Datenmaterial genau überlegt, wie stark leichte Veränderungen der zugrunde liegenden Variablen die gebildete Variable beeinflussen können. Ist der Einfluss sehr groß, steigt die Wahrscheinlichkeit, dass mit den synthetischen Daten keine plausiblen Ergebnisse erzielt werden können.

6 Neue Gewichte für synthetische Datensätze

Um mit den synthetischen Datensätzen valide Ergebnisse erzielen zu können, wurden von Infratest neue Gewichte bereitgestellt. Da eine gleichzeitige Anpassung sowohl an die Anzahl der Betriebe, als auch an die Beschäftigtenzahl, wie sie für das Originalbetriebspanel durchgeführt wird, einen nicht tragbaren Aufwand bedeutet hätte, wurden für die synthetischen Daten zwei Gewichte erzeugt. Ein Betriebsgewicht (gew_bet), das bei Auszählungen auf Betriebsebene zu verwenden ist, und ein Beschäftigtengewicht (gew_besch), das für Auszählungen auf Beschäftigtenebene verwendet werden kann. Diese Gewichte sind insbesondere zur Berechnung von Totalwerten gedacht. Bei der Berechnung von Mittelwerten (z.B. bei der Schätzung der durchschnittlichen Anzahl an Arbeitern pro Betrieb in Bayern) kann es allerdings aufgrund des vereinfachten Gewichtungsverfahrens zu Verzerrungen kommen. Wir empfehlen daher bei der Berechnung von Mittelwerten den Gewichtungsfaktor hr2007q zu verwenden.

Literatur

Abowd, J. M.; Stinson, M.; Benedetto, G. (2006): Final report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project. Tech. Rep., U.S. Census Bureau Longitudinal Employer-Household Dynamics Program.

Domingo-Ferrer, J.; Drechsler, J.; Polet-tini, S. (2009): Report on Synthetic Data Files. Tech. Rep., Eurostat.

Drechsler, J.; Bender, S.; Rässler, S. (2008): Comparing Fully and Partially Synthetic Data Sets for Statistical Disclosure Control in the German IAB Establishment Panel. In: Transactions on Data Privacy, Bd. 1, S. 105 – 130.

Drechsler, J.; Dundler, A.; Bender, S.; Rässler, S.; Zwick, T. (2008): A New Approach for Disclosure Control in the IAB Establishment Panel–Multiple Imputation for a Better Data Access. In: Advances in Statistical Analysis, Bd. 92, S. 439 – 458.

Graham, P.; Penny, R. (2005): Multiply Imputed Synthetic Data Files. Tech. Rep., University of Otago, http://www.uoc.otago.ac.nz/departments/pubhealth/pgrahpub.htm.

Graham, P.; Young, J.; Penny, R. (2009): Multiply Imputed Synthetic Data: Evaluation of Hierarchical Bayesian Imputation Models. In: Journal of Official Statistics, Bd. 25, S. 407–426.

Jacobebbinghaus, P.; Müller, D.; Orban, A. (2010): How to use data swapping to create useful dummy data for panel datasets. Tech. Rep., FDZ-Methodenreport, No. 3 (2010).

Little, R. J. A. (1993): Statistical Analysis of Masked Data. In: Journal of Official Statistics, Bd. 9, S. 407–426.

Machanavajjhala, Ashwin; Kifer, Daniel; Abowd, John M.; Gehrke, Johannes; Vilhuber, Lars (2008): Privacy: Theory meets Practice on the Map. In: ICDE, S. 277–286.

Reiter, J. P.; Raghunathan, T. E. (2007): The multiple adaptations of multiple imputation. In: Journal of the American Statistical Association, Bd. 102, S. 1462–1471.

Rubin, D. B. (1987): Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons.

Impressum

FDZ-Methodenreport 1/2011

Herausgeber
Forschungsdatenzentrum (FDZ)
der Bundesagentur für Arbeit
im Institut für Arbeitsmarkt- und Berufsforschung
Regensburger Str. 104
90478 Nürnberg

Redaktion
Stefan Bender, Dagmar Herrlinger

Technische Herstellung
Dagmar Herrlinger

Bezugsmöglichkeit
http://doku.iab.de/fdz/reporte/2011/MR_01-11.pdf

Internet
http://fdz.iab.de/

Rückfragen zum Inhalt an:
Jörg Drechsler,
Institut für Arbeitsmarkt und
Berufsforschung (IAB),
Weddigenstr. 20-22,
90478 Nürnberg
Telefon: 0911 / 179-4021
E-Mail: Joerg.Drechsler@iab.de