Dr. Jörg Drechsler

Erzeugung synthetischer Datensätze durch multiple Imputation: Theorie und Implementierung in der Praxis

Generating Multiply Imputed Synthetic Datasets: Theory and Implementation

Im November 2010 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. Ullrich Heilemann (Universität Leipzig), dem neuen Vorsitzenden des unabhängigen Gutachtergremiums, vorgetragenen Laudationes wurden in der Ausgabe 12/2010 dieser Zeitschrift veröffentlicht. In den Ausgaben Januar und Februar 2011 haben zwei der drei Preisträger des Jahres 2010 ihre Arbeiten in eigenen Beiträgen näher erläutert. Die Reihe über die im Jahr 2010 ausgezeichneten Arbeiten wird mit dem folgenden Beitrag von Herrn Dr. Jörg Drechsler abgeschlossen.

Die bei Frau Professor Dr. Susanne Rässler an der Otto-Friedrich-Universität Bamberg entstandene Dissertation "Generating Multiply Imputed Synthetic Datasets: Theory and Implementation" von Herrn Dr. Jörg Drechsler wurde mit dem Gerhard-Fürst-Preis 2010 in der Kategorie „Dissertationen" ausgezeichnet.

Vorbemerkung

Statistische Ämter wie das Statistische Bundesamt und die Statistischen Ämter der Länder wie auch verschiedene Forschungseinrichtungen tragen anhand von Registern und Befragungen viele wichtige statistische Informationen über Wirtschaft und Gesellschaft zusammen. Um eine effiziente Nutzung der erhobenen Daten zu fördern und eine breitgefächerte wissenschaftliche Forschung zu ermöglichen, wäre es wünschenswert, diese Informationen der interessierten Fachöffentlichkeit uneingeschränkt zur Verfügung stellen zu können. Ein uneingeschränkter Datenzugang könnte die wissenschaftliche Forschung stimulieren, Forschungsaufträge zu anstehenden politischen Entscheidungen auf eine solide Datengrundlage stellen, und Studierenden die Möglichkeit bieten, schon während ihres Studiums empirische Analysemethoden für komplexe Datensätze praktisch anzuwenden. Dies ist allerdings aus datenschutzrechtlichen Gründen nur in seltenen Fällen möglich.

Neben der gesetzlichen und ethischen Verpflichtung zum Datenschutz haben statistische Ämter und wissenschaftliche Institute aber auch ein starkes Eigeninteresse, die Anonymität der Befragten zu gewährleisten. Von der amtlichen Statistik wird die statistische Geheimhaltung als unverzichtbares Korrelat zu der bei den Erhebungen bestehenden Auskunftspflicht gesehen. Außerdem soll sie das Vertrauensverhältnis zwischen den Befragten und den statistischen Ämtern wahren. Auch bei Erhebungen der wissenschaftlichen Forschungsinstitute ist die Gewährleistung von Anonymität eine wichtige Voraussetzung, um Befragte für eine Teilnahme zu motivieren. Existieren Zweifel, ob der zugesicherte Datenschutz eingehalten wird, 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 diesen Fällen katastrophal sein.

Ein einfaches Entfernen identifizierender Merkmale wie Name und Anschrift ist oft nicht ausreichend, um die Anonymität der Befragungsteilnehmer zu gewährleisten. So zeigte zum Beispiel Sweeney[1], dass 97 % der Bürger in Cambridge,

1 Sweeney, L.: "Computational disclosure control for medical microdata: the Datafly system" in Proceedings of an International Workshop and Exposition, 1997, Seite 442 ff.

Massachusetts, die sich in Wählerlisten hatten registrieren lassen, allein über ihren Geburtstag und die Postleitzahl eindeutig identifiziert werden konnten. Mithilfe dieser öffentlich zugänglichen Wählerlisten konnte sie zum Beispiel den Gouverneur von Massachusetts in einer angeblich anonymisierten medizinischen Datenbank lokalisieren.

Um einen ausreichenden Datenschutz zu gewährleisten, setzen die statistischen Ämter daher verschiedene Anonymisierungsverfahren ein, bevor einzelne Datensätze dem interessierten Fachpublikum zugänglich gemacht werden. Häufig werden dabei Variablen vergröbert, indem beispielsweise statt des exakten Einkommens nur die Zugehörigkeit zu bestimmten Einkommensklassen weitergegeben wird. Auch detaillierte regionale Informationen werden oft auf Bezirks- oder Bundeslandsebene aggregiert. Daneben werden kontinuierliche Variablen häufig zensiert, indem zum Beispiel sehr hohe Einkommen als „100 000 Euro und mehr“ ausgewiesen werden. So wird beispielsweise im aktuellen Scientific-Use-File des Mikrozensus unter anderem das Einkommen in 25 Einkommensklassen angegeben, die Regionalinformationen werden auf Bundesländerebene vergröbert und das Alter der ältesten Befragungsteilnehmer wird mit „95 Jahre und älter“ ausgewiesen.[2] Diese Maßnahmen haben zwangsläufig einen Informationsverlust zur Folge. Durch die Aggregation können zum Beispiel keine kleinräumigen regionalen Auswertungen mehr durchgeführt werden und die Zensierung macht eine Analyse an den Rändern der betroffenen Verteilungen unmöglich. Oft sind aber gerade die Entwicklungen und Veränderungen an den Rändern einer Verteilung (beispielsweise für Personen mit besonders hohem beziehungsweise niedrigem Einkommen) von großer wirtschaftlicher oder politischer Bedeutung. Außerdem reichen die informationsreduzierenden Maßnahmen allein oft nicht aus, um eine vollständige Anonymität zu gewährleisten. Gerade bei Betriebsbefragungen genügen in der Regel sehr wenige Informationen, um einen Betrieb eindeutig zu identifizieren. Daher werden vor der Datenbereitstellung bei sensiblen Datensätzen zusätzlich sogenannte datenverändernde Verfahren eingesetzt, bei denen die Information auf Individualebene leicht verändert wird, um eine Re-Identifikation auszuschließen. Neben der Möglichkeit des “data swapping”, bei dem die Informationen zwischen einzelnen Befragungsteilnehmern ausgetauscht werden, wird häufig die sogenannte Mikro-Aggregation angewendet. Bei diesem Verfahren werden mehrere Beobachtungen mit hohem Identifikationsrisiko (zum Beispiel die Beschäftigtenzahl bei sehr großen Unternehmen) zusammengefasst und die einzelnen Werte zum Beispiel durch Mittelwerte ersetzt. Eine weitere Alternative wäre beispielsweise das Hinzufügen von Störtermen zu jeder Beobachtung einzelner Variablen (Noise Addition). Allerdings lassen sich die einzelnen Verfahren oft nur auf einen Teil der Variablen anwenden oder die Daten müssen so stark verändert werden, um einen ausreichenden Datenschutz zu gewährleisten, dass keine validen Analyseergebnisse erzielt werden können. So verzerrt das data swapping die Korrelation zwischen den betroffenen Variablen und solchen Variablen, bei denen keine Änderungen vorgenommen wurden. Die Mikro-Aggregation führt zu einer Unterschätzung der Varianz und bei der Überlagerung mit Störtermen entstehen Messfehler, die nur mit aufwendigen Korrekturschätzern wieder herausgerechnet werden können. Purdam und Elliot[3] zeigen anhand der allgemein zugänglichen Datensätze (Public-Use-Files) des Zensus im Vereinigten Königreich, wie stark die Auswirkungen verschiedener Anonymisierungsverfahren auf die Analyseergebnisse sein können, selbst wenn sie nur in geringem Umfang eingesetzt werden.

Ein sehr innovativer und in Europa noch relativ unbekannter Ansatz zur Anonymisierung, der die oben genannten negativen Auswirkungen vermeidet, ist die Erzeugung synthetischer Datensätze. Bei diesem von Rubin[4] erstmals vorgeschlagenen Verfahren, das auf den Ideen der multiplen Imputation beruht, werden die Originalwerte durch mehrere künstliche Werte ersetzt, die möglichst ähnliche Verteilungseigenschaften wie die Originaldaten aufweisen. Diese künstlichen oder synthetischen Daten werden dann der Allgemeinheit zur Verfügung gestellt. Potenzielle Nutzer können unter Verwendung einfacher Formeln für eine Vielzahl von Auswertungen valide Ergebnisse erzielen. Da es sich um rein fiktive Werte handelt, ist das Re-Identifikationsrisiko in der Regel zu vernachlässigen. Eine Vergröberung oder Zensierung ist somit auch nicht notwendig. Der Informationsgehalt bleibt also vollständig gewahrt.

Aufgrund seiner Anwendbarkeit auch für sehr komplexe zusammengespielte Paneldatensätze wird der Ansatz zur Erzeugung 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.[5] 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 Vereinigten Staaten sehr detailliert grafisch anzeigen lassen. Die zugrunde liegenden Daten wurden durch Erzeugung synthetischer Datensätze anonymisiert.[6] Weitere synthetische Datensätze sind derzeit in den Vereinigten Staaten in Entwicklung (The Longitudinal Business Database, The Longitudinal Employer-Household Dynamics Survey, The American Communities Survey).

In den letzten Jahren wurden in der Literatur verschiedene Varianten zur Erzeugung synthetischer Daten vorgeschlagen. Im Rahmen der hier vorgestellten Arbeit wurden diese

2 Statistisches Bundesamt, GESIS (Herausgeber): „Datenhandbuch zum Mikrozensus Scientific Use File 2008“, Wiesbaden 2010, im Internet unter *http://www.gesis.org/missy/fileadmin/missy/erhebung/datenhandbuch/DHB_2008.pdf* (abgerufen am 11. April 2011).

3 Purdam, K./Elliott, M.: “A case study of the impact of statistical disclosure control on data quality in the individual UK Samples of Anonymised Records” in Environment and Planning A, 2007, Jahrgang 39 (5), Seite 1101 ff.

4 Rubin, D. B.: “Discussion: Statistical disclosure limitation” in Journal of Official Statistics, Jahrgang 9, 1993, Seite 461 ff.

5 Siehe Abowd, J. M./Stinson, M./Benedetto, G.: “Final report to the social security administration on the SIPP/SSA/IRS public use file project”, Technical report, U.S. Census Bureau Longitudinal Employer-Household Dynamics Program, 2006.

6 Siehe Machanavajjhala, A./Kifer, D./Abowd, J. M./Gehrke, J./Vilhuber, L.: “Privacy: Theory meets practice on the map.” in ICDE 2008, Seite 277 ff.

Verfahren miteinander verglichen und jeweils auf das Betriebspanel des Instituts für Arbeitsmarkt- und Berufsforschung (IAB) der Bundesagentur für Arbeit angewendet. Ein wichtiges Ergebnis dieser Arbeit sind die synthetischen Datensätze der Welle 2007 des IAB-Betriebspanels, die seit Anfang 2011 über das Forschungsdatenzentrum der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- und Berufsforschung verfügbar sind.[7] Außerdem wird ein neues zweistufiges Imputationsverfahren vorgestellt, das eine bessere Abwägung zwischen der Begrenzung des Re-Identifikationsrisikos und einer möglichst hohen Datenqualität zulässt. Daneben werden neue Maße vorgeschlagen, um das verbleibende Re-Identifikationsrisiko der synthetischen Datensätze zu messen. Im Folgenden sollen die einzelnen Verfahrensvarianten und wichtige Ergebnisse der Arbeit kurz vorgestellt werden.

Erzeugung synthetischer Datensätze

Die ursprüngliche Idee zur Erzeugung synthetischer Datensätze geht zurück auf Rubin (siehe Fußnote 4). Das grundsätzliche Verfahren hat sich seitdem wenig geändert, auch wenn in den letzten Jahren verschiedene Verfahrensvarianten entwickelt wurden: Bayesianisch motiviert wird ein Modell an die Originaldaten angepasst, um eine A-posteriori-Verteilung, gegeben die beobachteten Daten, zu schätzen. Aus dieser Verteilung werden dann wiederholt plausible Werte gezogen. Im Gegensatz zum ursprünglichen Konzept der multiplen Imputation werden aber nicht für fehlende Beobachtungen plausible Werte generiert. Stattdessen werden alle Werte oder zumindest die Beobachtungen, die ein Re-Identifikationsrisiko darstellen, durch mehrere Züge aus der A-posteriori-Verteilung ersetzt. So entstehen für einen Originaldatensatz jeweils mehrere synthetische Datensätze. Ähnlich wie bei der Ergänzung fehlender Werte stellt die mehrfache Imputation sicher, dass die zusätzliche Varianz, die bei der Imputation entsteht, korrekt berücksichtigt werden kann. Ausgehend von m erzeugten synthetischen Datensätzen, erhält der Nutzer die finalen Ergebnisse durch die Anwendung einfacher Kombinationsregeln. Es sei $q^{(i)}$, $i = 1, \ldots, m$, der Punktschätzer (zum Beispiel der Mittelwert einer Variablen) im i-ten synthetischen Datensatz und $u^{(i)}$ der zugehörige Schätzer für die Varianz des Punktschätzers. Zur Berechnung der finalen Schätzer sind folgende drei Kenngrößen erforderlich:

$$(1) \qquad \bar{q}_m = \sum_i q^{(i)} / m$$

$$(2) \qquad \bar{u}_m = \sum_i u^{(i)} / m$$

$$(3) \qquad b_m = \sum_i (q^{(i)} - \bar{q}_m)^2 / (m-1)$$

Während sich der finale Punktschätzer $\bar{q}_m$ bei allen Verfahrensvarianten gleich berechnet, unterscheiden sich die Varianzschätzer zwischen den einzelnen Varianten, basieren jedoch immer auf einer Kombination der Varianzen innerhalb der Datensätze ($\bar{u}_m$) und zwischen den Datensätzen (b_m). Die hier vorgestellten Kombinationsregeln beziehen sich nur auf univariate Schätzer wie Mittelwerte oder Regressionskoeffizienten. Das korrekte Vorgehen bei multivariaten Schätzern, beispielsweise für multivariate Hypothesentests, ist in Reiter sowie Kinney und Reiter[8] beschrieben.

Vollständig synthetische Datensätze

Bei der Erzeugung vollständig synthetischer Datensätze werden grundsätzlich alle beobachteten Werte durch künstliche Werte ersetzt. In der Regel sind dazu zusätzliche Variablen nötig, die für die Grundgesamtheit vollständig 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.

Um valide Ergebnisse mit den Datensätzen zu erzielen, benötigt der Nutzer neben dem Punktschätzer $\bar{q}_m$ einen Schätzer für dessen Varianz. Dieser Varianzschätzer ergibt sich für vollständig synthetische Daten als $T_f = (1 + m^{-1}) b_m - \bar{u}_m$.[9]

Ein deutlicher Vorteil vollständig synthetischer Datensätze liegt darin, dass keinerlei 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. Ein weiterer bedeutender Vorteil 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. Im Rahmen der vorliegenden Arbeit durchgeführte Untersuchungen mit ausgewählten Variablen des IAB-Betriebspanels zeigen, dass valide Analyseergebnisse mit vollständig synthetischen Datensätzen möglich sind.[10]

Allerdings sind auch einige schwerwiegende Nachteile mit diesem Ansatz verbunden: Zum einen sollten 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 Auf-

7 Siehe Drechsler, J.: „Methodenreport: Synthetische Scientific-Use-Files der Welle 2007 des IAB-Betriebspanels“ in FDZ-Methodenreport 01/2011 (Ergänzung der Redaktion).

8 Siehe Reiter, J. P.: “Significance tests for multi-component estimands from multiply-imputed, synthetic microdata” in Journal of Statistical Planning and Inference, Jahrgang 131, 2005, Seite 365 ff., sowie Kinney, S. K./Reiter, J. P.: “Tests of Multivariate Hypotheses when using Multiple Imputation for Missing Data and Disclosure Limitation” in Journal of Official Statistics, Jahrgang 26, 2010, Seite 301 ff.

9 Siehe Raghunathan, T. E./Reiter, J. P./Rubin, D. B.: “Multiple imputation for statistical disclosure limitation” in Journal of Official Statistics, Jahrgang 19, 2003, Seite 1 ff.

10 Siehe Drechsler, J./Dundler, A./Bender, S./Rässler, S./Zwick, T.: “A new approach for disclosure control in the IAB Establishment Panel – Multiple imputation for a better data access” in Advances in Statistical Analysis, Jahrgang 92, 2008, Seite 439 ff.

wand 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 beziehungsweise 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 anderer Variablen verwendet werden kann. Basierend auf diesen Überlegungen scheint die Erzeugung teilweiser synthetischer Daten besonders attraktiv.

Teilweise synthetische Daten

Bei diesem von Little[11] vorgeschlagenen Verfahren werden keine neuen Stichproben künstlich generiert. Stattdessen werden 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 diesem Unternehmen zugeordnet werden können). Alle übrigen Variablen werden nicht verändert. Entsprechend umfasst der partiell synthetische Datensatz genau die Einheiten, die an der Befragung teilgenommen haben.

Die Kombinationsregeln für den Varianzschätzer bei teilweise synthetischen Daten unterscheiden sich geringfügig gegenüber dem Schätzer für vollständig synthetische Daten. Um valide Ergebnisse zu erzielen, muss der Schätzer $T_p = b_m/m + \bar{u}_m$ verwendet werden. Der Punktschätzer $\bar{q}_m$ entspricht dem Schätzer bei vollständig synthetischen Daten.[12]

Da bei teilweise synthetischen Datensätzen nur diejenigen Variablen imputiert werden, die ein Identifikationsrisiko bergen, kann die Abhängigkeit von der Qualität des gewählten Modells deutlich sinken, 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, sodass 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 zudem die Anzahl der sehr arbeitsaufwendigen Modellspezifikationen. Außerdem entfällt der Zwischenschritt der Generierung neuer Stichproben.

Auf der anderen Seite ist das verbleibende Re-Identifikationsrisiko 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 Identifikationsrisikos 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, damit eine eindeutige Re-Identifikation ausgeschlossen ist. Eine Möglichkeit zur Abschätzung des verbleibenden Risikos wird beispielsweise in Drechsler und Reiter[13] vorgestellt.

In der vorliegenden Arbeit wurden die Verfahren zur Erstellung teilweiser und vollständig synthetischer Datensätze vergleichend gegenübergestellt.[14] Im Rahmen einer empirischen Studie mit ausgewählten Variablen des IAB-Betriebspanels konnte gezeigt werden, dass mit beiden Verfahren Datensätze erzeugt werden können, die ein hohes Analysepotenzial aufweisen. Erwartungsgemäß fällt die Datenqualität der teilweise synthetischen Daten höher aus als die der vollständig synthetischen Daten. Die Untersuchungen zum Re-Identifikationsrisiko zeigen, dass auch mit den teilweise synthetischen Daten eine hohe Datensicherheit gewährleistet werden kann, sodass zumindest für diesen Datensatz die Erzeugung teilweise synthetischer Daten der Erzeugung vollständig synthetischer Daten vorzuziehen ist.

Multiple Imputation zur Ergänzung fehlender Werte und zur Erzeugung synthetischer Datensätze

Fehlende Werte stellen bei vielen Befragungen ein Problem dar. Da das Verfahren zur Erzeugung synthetischer Daten auf den Ideen der multiplen Imputation beruht, bietet es sich an, das Verfahren zuerst zur Imputation fehlender Werte zu verwenden und dann in einem zweiten Schritt die vollständig imputierten Datensätze zu synthetisieren. Weil sich die Modelle zur Imputation der fehlenden Werte von den Modellen zur Erzeugung synthetischer Datensätze unterscheiden können, ist es wichtig, die beiden Schritte unabhängig durchzuführen. Nach der Erstellung der m verschiedenen Datensätze, bei denen alle fehlenden Werte mehrfach imputiert wurden, müssen für jeden dieser Datensätze r synthetische Datensätze erzeugt werden. Die Gesamtzahl der erzeugten Datensätze ergibt sich somit als $M = mr$. Dieses zweistufige Verfahren bei der Erstellung der

11 Little, R. J. A.: "Statistical Analysis of Masked Data" in Journal of Official Statistics, Jahrgang 9, 1993, Seite 407 ff.

12 Siehe Reiter, J. P.: "Inference for partially synthetic, public use microdata sets" in Survey Methodology, Jahrgang 29, 2003, Seite 181 ff.

13 Drechsler, J./Reiter, J. P.: "Accounting for intruder uncertainty due to sampling when estimating identification disclosure risks in partially synthetic data" in Domingo-Ferrer, J./Saygin, Y. (Herausgeber): "Privacy in Statistical Databases", New York 2008, Seite 227 ff.

14 Siehe Drechsler, J./Bender, S./Rässler, S.: "Comparing Fully and Partially Synthetic Datasets for Statistical Disclosure Control in the German IAB Establishment Panel", Transactions on Data Privacy, 2008, Ausgabe 3, Seite 105 ff.

Daten muss auch bei der Analyse berücksichtigt werden. Es sei $q^{(i,j)}$, $i = 1, \ldots, m$, $j = 1, \ldots, r$, der Punktschätzer im j-ten synthetischen Datensatz des i-ten imputierten Datensatzes und $u^{(i,j)}$ die zugehörige Varianz des Punktschätzers. Zur Berechnung der Gesamtvarianz sind folgende Formeln notwendig:

$$(4) \qquad \bar{q}_r^{(i)} = \sum_j q^{(i,j)} / r$$

$$(5) \qquad \bar{q}_M = \sum_i \sum_j q^{(i,j)} / (mr)$$

$$(6) \qquad b_M = \sum_i (\bar{q}_r^{(i)} - \bar{q}_M)^2 / (m-1)$$

$$(7) \qquad \bar{w}_M = \sum_i \sum_j (q^{(i,j)} - \bar{q}_r^{(i)})^2 / m(r-1)$$

$$(8) \qquad \bar{u}_M = \sum_{i,j} u^{(i,j)} / (mr)$$

Der Punktschätzer $\bar{q}_M$ berechnet sich auch in diesem Fall als Mittelwert der Punktschätzer aus den einzelnen Datensätzen. Die Varianz des Punktschätzers berechnet sich als $T_M = (1 + m^{-1}) b_M - \bar{w}_M / r + \bar{u}_M$.

Im Rahmen der vorliegenden Arbeit wurde dieses zweistufige Verfahren verwendet, um für die Welle 2007 des IAB-Betriebspanels ein synthetisches Scientific-Use-File zu erzeugen. Dieser Datensatz konnte Anfang 2011 der interessierten Forschungsöffentlichkeit über das Forschungsdatenzentrum der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- und Berufsforschung zur Verfügung gestellt werden.[15] Der Bereitstellung gingen umfangreiche Analysen zum Re-Identifikationsrisiko voraus, mit denen nachgewiesen werden konnte, dass die Daten einen ausreichenden Datenschutz gewährleisten. Die durchgeführten Untersuchungen zum Analysepotenzial zeigen überwiegend sehr gute Ergebnisse. Die folgende Beispielanalyse soll die hohe Datenqualität veranschaulichen. Für weitere Ergebnisse der Untersuchungen und zur Veranschaulichung der Grenzen der Analysemöglichkeiten sei auf Drechsler[16] verwiesen.

Um einen fairen Vergleich zu gewährleisten, werden die Ergebnisse der Analysen mit den synthetischen Datensätzen stets mit denen der Analysen mit den Originaldaten des IAB-Betriebspanels nach der Imputation der fehlenden Werte verglichen. Neben der Gegenüberstellung der Punktschätzer wird auch ausgewiesen, wie stark sich die 95 %-Konfidenzintervalle der Punktschätzer überlappen.[17] Das Maß liegt immer zwischen 0 und 1, wobei die 0 keine Überlappung, die 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 ausgewiesenen 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 wurde für den Westen und den Osten Deutschlands unabhängig durchgeführt, aus Platzgründen werden hier aber nur die Resultate für den Westen präsentiert. Die Analysen für den Osten weisen ähnlich gute Ergebnisse auf.

Die Ergebnisse zeigen deutlich die hohe Datenqualität. Alle Punktschätzer liegen sehr nahe an den Punktschätzern der Originaldaten und die Überlappung des Konfidenzintervalls liegt bei mehr als 90 % für die meisten Koeffizienten. Auch die t-Werte liegen sehr nah an denen der Originaldaten, sodass die Analyse mit den synthetischen Daten die glei-

15 Im Internet unter *http://fdz.iab.de/de/FDZ_Establishment_Data/IAB_Establishment_Panel/IAB_Establishment_Panel_Outline/suf.aspx.*, abgerufen am 11. April 2011).

16 Drechsler, J.: "New Data Dissemination Approaches in old Europe – Synthetic Datasets for a German Establishment Survey", Journal of Applied Statistics (im Erscheinen).

17 Siehe Karr, A. F./Kohnen, C. N./Oganian, A./Reiter, J. P./Sanil, A. P.: "A framework for evaluating the utility of data altered to protect confidentiality" in The American Statistician, Jahrgang 60, 2006, Seite 224 ff.

Tabelle 1 Regressionsergebnisse einer Probit Regression von Teilzeitbeschäftigten (ja/nein) auf 19 erklärende Variablen im früheren Bundesgebiet

	Originaldaten	Synthetische Daten	Überlappung der 95 %-Konfidenzintervalle	*t*-Wert original	*t*-Wert synthetisch
Achsenabschnitt	– 0,809	– 0,752	0,87	– 7,23	– 6,85
5 bis 9 Beschäftigte	0,443	0,437	0,97	8,52	7,99
10 bis 19 Beschäftigte	0,658	0,636	0,90	11,03	10,88
20 bis 49 Beschäftigte	0,797	0,785	0,95	13,02	12,36
100 bis 199 Beschäftigte	0,892	0,908	0,96	9,23	9,48
200 bis 499 Beschäftigte	1,131	1,125	0,99	9,99	9,87
500 Beschäftigte und mehr	1,668	1,641	0,97	8,22	8,33
Beschäftigungswachstum erwartet	0,010	0,006	0,98	0,18	0,12
Beschäftigungsrückgang 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 (sechs Monate)	0,394	0,369	0,87	8,33	7,82
Entlassungen (sechs 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

chen Rückschlüsse zulässt wie die Analyse der Originaldaten. Weitere Untersuchungsergebnisse, die die hohe Datenqualität sowohl für multivariate Analysen als auch für eine Vielzahl deskriptiver Analysen unterstreichen, finden sich bei Drechsler (siehe Fußnote 16).

Ein zweistufiges Imputationsverfahren zur Optimierung des Trade-offs zwischen Datennutzen und Re-Identifikationsrisiko

Neben den konkreten Ergebnissen für das IAB-Betriebspanel wurden im Rahmen der Dissertation auch grundsätzliche Erkenntnisse gewonnen, die die Entwicklung und Analyse synthetischer Daten weiterführen. Es wurde ein zweistufiges Imputationsverfahren entwickelt, das es dem Datenproduzenten ermöglicht, besser zwischen einem hohen Analysepotenzial und einer hohen Datensicherheit zu vermitteln. Da bei der Erzeugung synthetischer Datensätze im Allgemeinen sowohl die Datenqualität als auch das Re-Identifikationsrisiko mit der Anzahl der erzeugten Datensätze steigen, erlaubt dieses Verfahren, verschiedene Variablen unterschiedlich oft zu imputieren. Dies hat den Vorteil, dass Variablen, die eine Vielzahl an Imputationen brauchen, um eine hohe Datenqualität zu gewährleisten, oft imputiert werden können. Variablen hingegen, die besonders für das Re-Identifikationsrisiko verantwortlich sind, werden nur wenige Male imputiert. Bei Anwendung dieses Verfahrens sind allerdings neue Schätzer für die Varianz nötig. Diese wurden in Reiter und Drechsler[18] hergeleitet. Unter Verwendung der Formeln (4) bis (6), wobei in diesem Fall m der Anzahl der Imputationen auf der ersten Stufe und r der Anzahl der Imputationen auf der zweiten Stufe entspricht, berechnet sich der Varianzschätzer für vollständig synthetische Daten als $T_{2st,f} = (1+m^{-1})b_M + (1-r^{-1})\overline{w}_M - \overline{u}_M$. Der Varianzschätzer für teilweise synthetische Daten ergibt sich als $T_{2st,p} = \overline{u}_M + b_M/m$.

Im Rahmen der Arbeit konnte durch empirische Analysen mit dem IAB-Betriebspanel gezeigt werden, dass durch das zweistufige Verfahren bei gleichbleibender Datenqualität eine deutliche Reduktion des Re-Identifikationsrisikos erreicht werden kann.[19]

Fazit

Die Erzeugung synthetischer Datensätze ist ein innovativer Ansatz, um den Datenzugang für externe Wissenschaftlerinnen und Wissenschaftler zu erweitern, ohne den zugesicherten Datenschutz zu gefährden. Gerade für besonders sensible Daten wie die aus Betriebsbefragungen, bei denen einfache informationsreduzierende Verfahren keine ausreichende Sicherheit bieten können, stellt das Verfahren eine interessante Alternative zu herkömmlichen datenverändernden Verfahren da. Im Rahmen der vorliegenden Dissertation konnte gezeigt werden, dass eine Erstellung synthetischer Datensätze auch für komplexe Datensätze wie die des IAB-Betriebspanels möglich ist. Mit der Freigabe der synthetischen Datensätze der Welle 2007 des IAB-Betriebspanels ist es von jetzt an möglich, diesen wichtigen Datensatz als Scientific-Use-File zu beziehen. Natürlich ist die Bereitstellung einer einzelnen Welle nur von begrenztem Nutzen. Die Attraktivität des Betriebspanels liegt in erster Linie in der Möglichkeit, Panelanalysen durchzuführen. Die vorgelegte Arbeit ist insofern als Machbarkeitsstudie zu verstehen, die einen ersten Schritt in Richtung einer vollständigen Anonymisierung des IAB-Betriebspanels darstellt. Natürlich können die bereitgestellten Datensätze nicht für jede vorstellbare Analyse valide Ergebnisse liefern. Ein Anonymisierungsverfahren, das eine vollständige Datensicherheit bei gleichzeitiger uneingeschränkter Analysefähigkeit bietet, kann es nicht geben. Insofern sollten die Datensätze auch lediglich als eine Erweiterung des bisherigen Datenangebots neben dem Datenfernrechnen und der Möglichkeit eines Gastaufenthalts am Forschungsdatenzentrum der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- und Berufsforschung verstanden werden. Bestimmte Analysen werden weiterhin nur mit den Originaldaten brauchbare Ergebnisse erzielen, denn nur die Zusammenhänge, die in den Imputationsmodellen zur Erstellung der synthetischen Daten berücksichtigt wurden, werden sich auch in den erzeugten Daten finden lassen. Wurden wichtige Zusammenhänge bei der Modellierung vernachlässigt, liefern die synthetischen Datensätze verzerrte Ergebnisse. Um dem Datennutzer eine Einschätzung zu ermöglichen, ob seine Analysen auf den bereitgestellten Datensätzen durchführbar sind, stehen umfangreiche Metadaten zur Verfügung, in denen für jede synthetisierte Variable aufgelistet ist, welche Zusammenhänge bei der Modellierung berücksichtigt wurden. Anhand dieser Liste kann der Nutzer dann entscheiden, ob die Daten für seine Analysen ausreichen oder ob er besser einen Aufenthalt im Forschungsdatenzentrum beantragt, um mit den Originaldaten rechnen zu können. Zusätzlich wird jedem Nutzer bis auf Weiteres zugesichert, dass die erstellten Analyseprogramme auch auf den Originaldaten gerechnet werden können und die Ergebnisse dem Nutzer nach einer Prüfung auf Einhaltung des Datenschutzes übermittelt werden. Somit erhält der Nutzer die Garantie, dass er am Ende immer die Originalergebnisse erhält.

Ob sich das vorgestellte Verfahren auf lange Sicht durchsetzen wird, hängt in erster Linie von der Akzeptanz durch die Nutzer ab. Bisher steht interessierten Nutzern weltweit nur eine sehr geringe Anzahl synthetischer Datensätze zur Verfügung. Allerdings werden in den nächsten Jahren vor allem in den Vereinigten Staaten verschiedene Datensätze als synthetische Datensätze der Allgemeinheit zugänglich gemacht werden. Dann wird sich erstmals beurteilen lassen, ob dieses Verfahren eine breite Zustimmung findet. In der Zwischenzeit unterstreichen die positiven Ergebnisse dieser Arbeit und die Resonanz, die die ersten in den Vereinigten Staaten verfügbaren Datensätze gefunden haben, dass dieser innovative Versuch das Potenzial hat, das immerwährende Spannungsverhältnis zwischen größtmöglichem Datenzugang und optimalem Datenschutz zu lösen.

18 Reiter, J. P./Drechsler, J.: „Releasing multiply-imputed synthetic data generated in two stages to protect confidentiality" in Statistica Sinica, Jahrgang 20, 2010, Seite 405 ff.

19 Siehe Drechsler, J./Reiter, J. P.: "Disclosure risk and data utility for partially synthetic data: An empirical study using the German IAB Establishment Survey", Journal of Official Statistics, Jahrgang 25, 2009, Seite 589 ff.

Auszug aus Wirtschaft und Statistik

Herausgeber
Statistisches Bundesamt, Wiesbaden

www.destatis.de

Schriftleitung
Roderich Egeler, Präsident des Statistischen Bundesamtes

Brigitte Reimann (verantwortlich für den Inhalt)
Telefon: +49 (0) 6 11 / 75 20 86

Ihr Kontakt zu uns
www.destatis.de/kontakt

Statistischer Informationsservice
Telefon: +49 (0) 6 11 / 75 24 05
Telefax: +49 (0) 6 11 / 75 33 30

Abkürzungen

WiSta	=	Wirtschaft und Statistik
MD	=	Monatsdurchschnitt
VjD	=	Vierteljahresdurchschnitt
HjD	=	Halbjahresdurchschnitt
JD	=	Jahresdurchschnitt
D	=	Durchschnitt (bei nicht addierfähigen Größen)
Vj	=	Vierteljahr
Hj	=	Halbjahr
a. n. g.	=	anderweitig nicht genannt
o. a. S.	=	ohne ausgeprägten Schwerpunkt
St	=	Stück
Mill.	=	Million
Mrd.	=	Milliarde

Zeichenerklärung

p	=	vorläufige Zahl
r	=	berichtigte Zahl
s	=	geschätzte Zahl
–	=	nichts vorhanden
0	=	weniger als die Hälfte von 1 in der letzten besetzten Stelle, jedoch mehr als nichts
.	=	Zahlenwert unbekannt oder geheim zu halten
. . .	=	Angabe fällt später an
X	=	Tabellenfach gesperrt, weil Aussage nicht sinnvoll
I oder —	=	grundsätzliche Änderung innerhalb einer Reihe, die den zeitlichen Vergleich beeinträchtigt
/	=	keine Angaben, da Zahlenwert nicht sicher genug
( )	=	Aussagewert eingeschränkt, da der Zahlenwert statistisch relativ unsicher ist

Abweichungen in den Summen ergeben sich durch Runden der Zahlen.

© Statistisches Bundesamt, Wiesbaden 2011

Vervielfältigung und Verbreitung, auch auszugsweise, mit Quellenangabe gestattet.