Erzeugung synthetischer Datensätze mittels Methoden der multiplen Imputation

Hans-Peter Hafner
Hessisches Statistisches Landesamt
Rheinstr. 35/37
65185 Wiesbaden
hhafner@statistik-hessen.de

Rainer Lenz
Hochschule für Technik und Wirtschaft des Saarlandes
Goebenstr. 40
66117 Saarbrücken
lenz@htw-saarland.de

Zusammenfassung

Nach §16 (6) des Bundesstatistikgesetzes haben wissenschaftliche Einrichtungen Zugang zu so genannten faktisch anonymisierten Einzeldaten der amtlichen Statistik. Danach müssen die Daten von den statistischen Ämtern soweit verändert werden, dass der mit einer Enthüllung eines vertraulichen Einzelwertes verbundene Aufwand eines potentiellen Datenangreifers den Nutzen übertrifft. Der von den Wissenschaftlern favorisierte Datenzugangsweg ist die so genannte kontrollierte Datenfernverarbeitung (KDFV), bei der sie dem statistischen Amt ein Programm zusenden, das dort an den Originaldaten ausgeführt wird. Bisher ist dieses Verfahren aber sowohl für den Wissenschaftler, als auch für die Mitarbeiter der statistischen Ämter sehr zeitaufwändig. Im Rahmen eines neuen Projektes sollen daher Datenstrukturfiles entwickelt werden, d. h. absolut anonyme Dateien, die aber in der Struktur den zugrunde liegenden Originaldaten möglichst ähnlich sind und mit denen die Wissenschaftler die Programme für die KDFV entwickeln und testen können, bevor sie später auf die Originaldaten angewendet werden.

Der vorliegende Beitrag befasst sich mit der Erstellung der anonymisierten Dateien. Dabei kommen erstmals in Deutschland für Daten der amtlichen Statistik Methoden der multiplen Imputation zum Einsatz. Für die Generierung dieser synthetischen Datensätze wird die Software IVEware verwendet, die sowohl als Stand-Alone-Version als auch als SAS – Implementierung verfügbar ist. Die Software wird vorgestellt und auf Betriebsdaten angewendet. Als erfolgreich kann eine Anonymisierungsmethode angesehen werden, wenn sie beide relevanten Ziele erreicht. Diese sind zum einen der bestmögliche Erhalt an Analysepotential in den Daten, und zum anderen eine ausreichende Minimierung des Risikos für eine mögliche Reidentifikation von Betrieben. Bei der Berechnung dieses Reidentifikationsrisikos wird von dem Szenario ausgegangen, dass ein potentieller Datenangreifer aus kommerziellen Datenbanken über Betriebsinformationen verfügt und damit versucht, in den anonymisierten Daten möglichst viele dieser Betriebe mittels der Verwendung von Ähnlichkeitsmaßen wiederzufinden. Die programmtechnische Umsetzung dieses Szenarios erfolgt im Wesentlichen mit SAS/IML.

Schlüsselwörter: Forschungsdatenzentren, synthetische Datensätze, Datenstrukturfiles, Reidentifikationsrisiko

1 Die Forschungsdatenzentren der amtlichen Statistik

Die Einrichtung der Forschungsdatenzentren (FDZ) geht auf die Empfehlungen der „Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik" zurück, die vom Bundesministerium für Bildung und Forschung eingerichtet wurde. In ihrem Gutachten „Wege zu einer besseren informationellen Infrastruktur“ aus dem Jahr 2001 stellt die Kommission fest, dass die Leistungsfähigkeit der Dateninfrastruktur eine entscheidende Grundlage für die Leistungsfähigkeit der Gesellschaft sowie für eine im internationalen Maßstab innovationsfähige sozial- und wirtschaftswissenschaftliche Forschung ist[1]. Daher müssen Daten – auch solche, die für andere Zwecke im Rahmen staatlichen Handelns entstehen – so effizient wie möglich für wissenschaftliche Analysen genutzt werden. Die Empfehlungen der Kommission wurden im Jahr 2001 vom Gründungsausschuss des Rates für Sozial- und Wirtschaftsdaten aufgegriffen, indem er die großen öffentlichen Datenproduzenten – darunter die statistischen Ämter – aufforderte, Forschungsdatenzentren einzurichten. Die Statistischen Ämter des Bundes und der Länder sind dieser Aufforderung nachgekommen. Das Forschungsdatenzentrum des Statistischen Bundesamtes wurde im Oktober 2001, das der Statistischen Landesämter im April 2002 eingerichtet.

Grundsätzlich darf die amtliche Statistik Einzelangaben nur dann zur Verfügung stellen, wenn diese absolut anonymisiert sind, d. h. eine Reidentifikation nicht möglich ist. Für die Wissenschaft existiert seit 1987 eine Ausnahmeregelung, die von diesem allgemeinen Grundsatz abweicht (siehe § 16 Abs. 6 Bundesstatistikgesetz (BStatG)). Demnach dürfen die statistischen Ämter Einzelangaben an Hochschulen und sonstige Einrichtungen mit der Aufgabe unabhängiger wissenschaftlicher Forschung übermitteln, die eine Deanonymisierung zwar nicht mit absoluter Sicherheit ausschließen, aber Betroffenen nur dann zugeordnet werden können, wenn der Datenempfänger für eine Zuordnung einen unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft erbringen muss. Dieses so genannte „Wissenschaftsprivileg“ ist Voraussetzung für die Nutzung der als ***faktisch anonym*** bezeichneten Mikrodaten, die der Wissenschaft für ein definiertes Forschungsvorhaben bereitgestellt werden. Die faktische Anonymität muss für jede einzelne Statistik unter Berücksichtigung des potentiell vorhandenen Zusatzwissens und möglicher Angriffsszenarien sichergestellt werden.

Bei den Datenzugangswegen lassen sich prinzipiell die Onsite- (im statistischen Amt) und die Offsite – Nutzung (am eigenen Arbeitsplatz) unterscheiden. Die Wissenschaftler können schwach anonymisierte Daten an den Gastwissenschaftlerarbeitsplätzen in den statistischen Ämtern nutzen und sie haben Zugriff auf formal anonymisierte Daten im Rahmen der so genannten kontrollierten Datenfernverarbeitung (KDFV). Bei der KDFV senden sie einem FDZ-Mitarbeiter ein in SAS, SPSS oder STATA geschriebenes Auswertungsprogramm, welches dieser an den Daten ausführt. Nach Prüfung der Ergebnisse auf Datenschutzbestimmungen, werden sie dem Wissenschaftler zugesandt.

[1] Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik (Hrsg.): Wege zu einer besseren informationellen Infrastruktur. Baden-Baden, S. 15.

Am eigenen Arbeitsplatz können die Forscher mit stärker anonymisierten Scientific Use Files (SUFs) arbeiten, während die sehr stark in ihrem Analysepotential reduzierten CAMPUS – Files speziell für die Lehre entwickelt werden.

2 Das Projekt InfinitE

Die Datenproduzenten beobachten einen fundamentalen Wandel in der Nachfrage nach Produkten der Wirtschaftsstatistiken. Bis vor einigen Jahren wurde die Bereitstellung von SUFs als der Königsweg angesehen für einen adäquaten Zugang der Sozial- und Wirtschaftswissenschaften zu amtlichen Mikrodaten in Deutschland. Solche SUFs sind verfügbar für ausgewählte und stark nachgefragte Statistiken. Jedoch werden SUFs für Wirtschaftsstatistiken nur bedingt angenommen. Ein Grund dafür sind die neuen datenverändernden Anonymisierungsmethoden, die angewandt werden müssen, um die Vertraulichkeit der Daten zu gewährleisten. Diese zerstören jedoch einen Teil der Struktur der Daten. Außerdem verstreicht im Allgemeinen zu viel Zeit zwischen der Datenerhebung und der Erstellung des zugehörigen SUF. In den letzten Jahren wurden die Nutzung der kontrollierten Datenfernverarbeitung und der Gastwissenschaftlerarbeitsplätze die am häufigsten genutzten Zugangswege zu den Daten der Wirtschaftsstatistiken in den FDZ der amtlichen Statistik. Deshalb ist das Ziel des im Juni 2009 gestarteten Projektes InfinitE (Eine informationelle Infrastruktur für das ‚E-Science-Age', siehe Lenz 2009) einerseits, anonymisierte Datenstrukturfiles zu entwickeln, die dazu verwendet werden können, ökonometrische Modelle zu spezifizieren und syntaktisch fehlerfreie Programm-Codes zu schreiben. Weiterhin soll die Kontrolle der Programmausgaben (Output), die für die Mitarbeiter der FDZ in der Regel äußerst zeitaufwändig ist, so weit wie möglich automatisiert werden.

3 Datenstrukturfiles

Datenstrukturfiles sind absolut anonymisierte Testdatensätze, die die gleiche Struktur besitzen wie die formal anonymisierten, also um direkte Identifikatoren wie etwa Name und Adresse der Einheiten gekürzte Originaldaten, auf welche die Wissenschaftler im Rahmen der KDFV zugreifen können. Die Datenstrukturfiles dienen zur Entwicklung der Analyseprogramme. Bisher bestehen die Datenstrukturfiles oftmals aus einer Stichprobe des Originalmaterials, auf welche zusätzlich starke Anonymisierungsmaßnahmen angewendet werden, oder es werden zufällig Einzelwerte aus dem vorgegebenen Wertebereich des zu verändernden Merkmals generiert. Bei beiden Vorgehensweisen bleiben die Merkmale zwar in den Daten erhalten, ihre Ausprägungen und die Abhängigkeitsstruktur (Filterführung der Fragen, Varianz-Kovarianz-Matrix etc.) zwischen den Merkmalen werden dabei aber komplett zerstört. Somit kann ein Forscher zwar prüfen, ob sein Programm lauffähig ist, bekommt aber keine Hinweise, ob die inhaltliche Fragestellung adäquat umgesetzt wurde. Daher können oftmals die Auswertungsprogramme der Wissenschaftler nicht eins zu eins für die spätere Anwendung auf die Originaldaten übernommen werden. Fast immer sind vor der endgültigen Anwendung auf die formal

anonymisierten Daten mehr oder weniger zeitaufwändige Anpassungsarbeiten durch die Wissenschaftler und FDZ-Mitarbeiter nötig.

3.1 Synthetische Datenstruktursätze

Eine Möglichkeit, Datenstrukturfiles mit deutlicher höherer Qualität bereitzustellen, bietet die Erzeugung synthetischer Datensätze basierend auf den Ideen zur multiplen Imputation fehlender Werte. Der entscheidende Vorteil bei diesem Verfahren liegt in der Universalität des Ansatzes. Sämtliche Restriktionen und Filterstrukturen können bei der Erstellung berücksichtigt werden, außerdem ist der Ansatz auf kontinuierliche Variablen ebenso anwendbar wie auf kategoriale Variablen. Aufgrund seiner hohen Flexibilität und Anwendbarkeit auch für sehr komplexe zusammen gespielte Paneldatensätze wird der innovative Ansatz in den letzten Jahren international immer stärker eingesetzt.
Der Vorschlag, mittels multipler Imputation erzeugte synthetische Datensätze für die Wissenschaft zu veröffentlichen, wurde erstmals in Rubin (1993) unterbreitet und in Raghunathan, Reiter und Rubin (2003) weiter ausgeführt. Das grundlegende Prinzip ist, jeweils mehrere synthetische Datensätze zu erzeugen, die einzeln analysiert werden. Das tatsächliche Analyseergebnis ergibt sich dann durch die Anwendung einfacher Kombinationsregeln (Raghunathan et al. (2001)).
Prinzipiell lassen sich voll synthetische und partiell synthetische Datensätze unterscheiden. Bei den voll synthetischen Datensätzen werden alle Einheiten der Grundgesamtheit, die nicht zur Stichprobe der Erhebung gehören, als fehlende Werte behandelt. Für diese Einheiten benötigt man aus der Grundgesamtheit Informationen (z. B. aus dem Unternehmensregister oder der Beschäftigtenstatistik), die dann in das Imputationsmodell einfließen. Die fehlenden Werte werden aus der posterioren Verteilung, gegeben die beobachteten Werte, imputiert. Verschiedene Stichproben dieser imputierten Werte werden dann der Wissenschaft zugänglich gemacht. Im Unterschied dazu werden bei den partiell synthetischen Datensätzen für die in der Erhebung enthaltenen Einheiten besonders sensible Merkmale und/oder Überschneidungsmerkmale zu öffentlich zugänglichen Datensätzen durch synthetische Werte ersetzt (s. dazu etwa Reiter 2003).

4 Erzeugung synthetischer Datensätze aus amtlichen Mikrodaten

Im genannten Projekt InfinitE sollen unterschiedliche Anonymisierungsstrategien zunächst beispielhaft für die Monatsberichte für Betriebe des verarbeitenden Gewerbes für die Wellen 1999 bis 2002 entwickelt und miteinander verglichen werden. Diese Erhebung wird einerseits von der Wissenschaft stark nachgefragt, andererseits besitzt sie einen überschaubaren Katalog an rund 30 Merkmalen.

4.1 Die Monatsberichte im verarbeitenden Gewerbe

Berichtspflichtig sind alle Betriebe, die ihren wirtschaftlichen Schwerpunkt im verarbeitenden Gewerbe haben und mindestens 20 Arbeitnehmer beschäftigen. Ebenfalls in der Erhebung enthalten sind kleinere Betriebe, falls das Unternehmen, zu dem er gehört, mindestens 20 Beschäftigte besitzt. Unter den Merkmalen, die berichtet werden, befinden sich der Wirtschaftszweig, der Betriebssitz (Gemeinde), die Anzahl der Beschäftigten, der Umsatz, gezahlte Löhne und Gehälter sowie geleistete Arbeitsstunden.

Wie der Name der Erhebung schon aussagt, werden die Daten monatlich erhoben. In den Forschungsdatenzentren sind zum jetzigen Zeitpunkt jedoch nur die aufsummierten Jahreswerte für wissenschaftliche Analysen verfügbar.

Im Folgenden stellen wir eine Software zur Erstellung synthetischer Datensätze genauer vor und wenden diese auf die Daten der Monatsberichte 2001 an.

4.2 Das Programm IVEWare

IVEware ist ein von Raghunathan, Solenberger und Van Hoewyk entwickeltes Programm, das u. a. die (multiple) Imputation fehlender Werte durchführt.[2] Dabei wird das Verfahren der sequentiellen Regression verwendet. D. h.: Seien $X_1, \ldots, X_i$ die Variablen des Datensatzes ohne fehlende Werte, $Y_1, \ldots, Y_j$ die Variablen des Datensatzes mit fehlenden Werten. Dabei sei die Anordnung der Y-Variablen nach der Anzahl der fehlenden Werte von den wenigsten bis zu den meisten. Im ersten Schritt wird ein Modell für die bedingte Verteilung von Y_1, gegeben die beobachteten X-Werte, geschätzt. Aus dieser Verteilung werden Werte für Y_1 imputiert. Im nächsten Schritt wird ein Modell für die bedingte Verteilung von Y_2 gegeben die beobachteten X-Werte und die imputierten Y_1–Werte geschätzt und aus dieser Verteilung werden Werte für Y_2 imputiert usw.[3] IVEware unterscheidet vier Typen von Variablen: stetige, kategoriale, gemischte (0 als kategorialer Wert, sonst stetig) und Zählvariablen. Für stetige Variablen wird das normale lineare Regressionsmodell zur Schätzung verwendet, für kategoriale ein logistisches oder verallgemeinertes logistisches Modell. Gemischte Variablen werden zweistufig imputiert: Zunächst wird mit einem logistischen Regressionsmodell Null oder Nicht–Null imputiert; für die Nicht–Null–Werte erfolgt dann die Imputation des Wertes mit einem linearen Regressionsmodell. Zählvariablen werden mit einer Poisson–Regression imputiert.

Seit Ende 2009 existiert eine neue Version von IVEware als Prototyp, die zum Zeitpunkt dieser Veröffentlichung noch nicht über das Internet erhältlich ist. Neu hinzugekommen ist u. a. der Befehl SYNTHESIZE, der die einfache Erstellung synthetischer Datensätze ermöglicht, ohne zuvor künstlich fehlende Werte erzeugen zu müssen. Weiterhin wurde die Behandlung kategorialer Merkmale komplett geändert. Während bisher

[2] Software und User-Guide können unter http://www.isr.umich.edu/src/smp/ive/ kostenlos herunter geladen werden

[3] Eine genauere Beschreibung des Verfahrens findet sich in Raghunathan et al. (2001)

für kategoriale Merkmale das multinomiale Modell zugrunde gelegt wurde, das für seine schlechten Konvergenzeigenschaften bekannt ist, erfolgt nun eine Zerlegung in Dummy-Variablen.

IVEware bietet prinzipiell zwei Möglichkeiten, um die Struktur der Originaldaten und Abhängigkeiten zwischen den Variablen zu erhalten. Mit dem Bounds–Statement lassen sich Ober- und Untergrenzen angeben, mit dem Restrict–Statement lässt sich angeben, dass eine Imputation nur durchgeführt werden soll, wenn bestimmte Bedingungen erfüllt sind, wie z. B. bei Personendaten eine Imputation der Anzahl von Geburten nur für weibliche Personen eines bestimmten Mindestalters.

IVEware kann innerhalb von SAS oder als unabhängiges Programm ausgeführt werden. Unter SAS läuft das Programm allerdings nicht unter der komfortablen Enterprise Guide–Oberfläche.

4.3 Entwicklung synthetischer Datensätze aus den Monatsberichten

Ein typischer Code für die Erzeugung synthetischer Datensätze sieht wie das nachstehende Beispiel aus:

```
%impute(name = testneu, dir = 'F:\_B Dateien und Dokumente\4
Projekte zur Datenverknüpfung\1
AFiD\Projekt_Monatsberichte_GLS\Betriebspanel', setup = new);
Datain mb.mb_gb_2001_aufbereitet_original;
Dataout mb.daten_synthetisch CON;
Default transfer;
Categorical m03_2001 m07_2001 m05_2001 wz2_2001;
Count m04_2001;
Continuous tm11_2001;
Mixed tm12_2001 tm13_2001 tm20_2001 tm21_2001 tm22_2001 tm24_2001
tm25_2001 tm31_2001 tm32_2001;
SYNTHESIZE m03_2001 m05_2001 wz2_2001 tm11_2001 tm12_2001 tm13_2001
tm20_2001 tm21_2001 tm22_2001 tm24_2001 tm25_2001 tm31_2001
tm32_2001 tm35_2001
tm36_2001 tm37_2001;
bounds m04_2001 (>= 1, <= 25) tm11_2001 (>= 1, < 40)
tm12_2001 (>= 0, < 10) tm13_2001 (>= 0, < 20)
tm16_2001 (>= 0, < 32) tm17_2001 (>= 0, < 7) tm18_2001 (>= 0, < 14)
tm20_2001 (>= 0, <= 1800) tm21_2001 (>= 0, < 430)
tm22_2001 (>= 0, < 2200) tm24_2001 (>= 0, < 2600) tm25_2001 (>= 0, <
2500) tm31_2001 (>= 0, < 2000) tm32_2001 (>= 0, < 2900)
tm35_2001 (>= 0, < 400) tm36_2001 (>= 0, < 1100) tm37_2001 (>= 0, <
1100);
Implicates 5;
Multiples 1;
Iterations 10;
run;
```

Der Befehl *Default* gibt den Typ von Merkmalen an, die im Code nicht eigens aufgeführt werden. *Default transfer* bedeutet, dass diese Merkmale unverändert in die synthetische Datei übernommen werden und dass sie darüber hinaus nicht in die Regressionsmodelle zur Erzeugung der imputierten bzw. synthetischen Daten eingehen. Mit *implicates* lässt sich die Anzahl der synthetischen Datensätze angeben, während *multiples* die Anzahl der Imputationen der fehlenden Werte vorgibt. Die Erzeugung der synthetischen Daten erfolgt zweistufig: Zunächst werden evtl. im Datensatz vorhandene fehlende Werte imputiert, danach werden auf der Grundlage der vollständigen lückenlosen Daten synthetische Daten generiert.

Insgesamt werden fünf synthetische Datensätze für die Monatsberichte erzeugt. Die stetigen Merkmale haben wir zuvor mit der dritten Wurzel transformiert (um eine Annäherung an eine Normalverteilung zu erreichen) und anschließend wieder rücktransformiert. In nachfolgender Tabelle werden Kennwerte bzgl. der Anzahl der tätigen Personen und des Gesamtumsatzes der synthetischen Daten denselben Größen der Originaldaten gegenüber gestellt.

Tabelle 1: Mittelwert und Standardabweichung für tätige Personen und Gesamtumsatz

Datensatz	Tätige Personen		Gesamtumsatz	
	Mittelwert	Standard-abweichung	Mittelwert	Standard-abweichung
Synthetisch 1	127,0	585,8	24494649	225562290
Synthetisch 2	127,2	583,0	24456424	216807664
Synthetisch 3	126,9	577,0	24502221	216667023
Synthetisch 4	127,3	576,4	24321639	208523554
Synthetisch 5	127,0	584,9	24355994	220182716
Original	128,2	575,9	26911974	274285595

Der Mittelwert der Anzahl der tätigen Personen liegt in den synthetischen Datensätzen um 0,7 bis 1 % unter dem originalen Durchschnittswert, während die Standardabweichungen leicht über dem wahren Wert liegen (maximal 1,7 %). Für den Umsatz sind die relativen Abweichungen größer: Beim Mittelwert erreichen sie bis zu 10, bei der Standardabweichung gar bis zu 24 %. Dies liegt daran, dass es beim Merkmal Umsatz recht viele Extremwerte gibt, für die die Anpassung des Regressionsmodells schlecht ist. Eine Verbesserung lässt sich unter Umständen mit einer besser geeigneten Transformationsfunktion erreichen. Würde man allerdings den natürlichen Logarithmus anstelle der dritten Wurzel wählen, würde dies zu einer Überschätzung der originalen Werte führen, die sich in derselben Größenordnung wie die derzeitige Unterschätzung bewegt. Da bei der Anonymisierung auch immer zu berücksichtigen ist, dass die Anpassung nicht zu gut sein darf, um gerade bei großen Werten und Ausreißern das Reidentifikationsrisiko in Grenzen zu halten, kann sich eine Über- oder Unterschätzung später noch als nützlich erweisen. Auf die Bestimmung des Reidentifikationsrisikos wird in Kapitel 5 näher eingegangen.

5 Simulation von Datenangriffen zwecks Ermittlung des Reidentifikationsriskos

Bei der Betrachtung vertraulicher Mikrodaten sind stets zwei gleichrangige Ziele zu verfolgen. Diese sind zum einen der bestmögliche Erhalt des Analysepotentials und zum anderen die Gewährleistung eines ausreichenden Grades an Anonymität. Letzterer kann via Simulation sogenannter Massenfischzüge festgestellt werden.

5.1 Mathematische Modellierung

Bei einem sogenannten Massenfischzug versucht ein potentieller Datenangreifer, seine möglicherweise kommerziell erworbene externe Datenquelle A mit den vertraulichen Zieldaten B zusammen zu führen. Dabei werden paarweise Informationen zweier Merkmalsträger $a \in A$ und $b \in B$ zusammengeführt, die als zu demselben Individuum gehörig vermutet werden. Im Folgenden wird das naheliegende Ziel verfolgt, die Anzahl der Fehlzuordnungen zu minimieren. Wenn wir dem Datenangreifer Kenntnis über die Teilnahme der gesuchten Unternehmen an der Zielerhebung unterstellen, kann dieses Problem der Zuordnung mathematisch wie folgt formuliert werden: Finde eine injektive Abbildung $\varphi : A \to B$, basierend auf dem Distanzmaß $d : A \times B \to [0,1]$ (oder alternativ dem Ähnlichkeitsmaß $w : A \times B \to [0,1]$), welche jeden Merkmalsträger in A auf den nächsten (oder ähnlichsten) Merkmalsträger in B abbildet. Genauer kann diese Abbildung über folgendes parametrisches lineares Zuordnungsproblem definiert werden:

$$\text{Minimiere} \quad \sum_{i=1}^{n} \sum_{j=1}^{n} d(a_i, b_j) x_{ij}, \qquad \textbf{(LP)}$$

unter $\quad x_{ij} \in \{0,1\} \quad \text{für} \quad i, j = 1,...,n,$

$$\sum_{j=1}^{n} x_{ij} = 1 \quad \text{für} \quad i = 1,...,n \qquad \text{und}$$

$$\sum_{i=1}^{n} x_{ij} = 1 \quad \text{für} \quad j = 1,...,n.$$

Die Nebenbedingungen von *(LP)* stellen sicher, dass jedem Merkmalsträger a der externen Daten A genau ein Merkmalsträger b der Zieldaten B zugeordnet wird. D. h., es gilt $x_{ij}=1$ genau dann, wenn a_i mit b_j verbunden ist. Daher erscheint es sinnvoll anzunehmen, dass A weniger als oder gleich viele Elemente wie B aufweist.

5.2 Lösung des Zuordnungsproblems mittels SAS/IML

Wir beginnen mit dem Auszug des SAS/IML-Codes zur Berechnung der standardisierten Distanzen:

```
D=j(n_a,n_b,0);
do i=1 to n_a;
   diff=A[,i];
     do j=2 to n_b;
          diff=diff||A[,i];
     end;
   diff=abs(diff-B);
   diff_min=T(diff[,><]);
   diff_max=T(diff[,<>]);
   diff_dupl_min=diff_min;
   diff_dupl_max=diff_max;
   if n_b > 1 then do k=2 to n_b;
             diff_dupl_min=diff_dupl_min//diff_min;
             diff_dupl_max=diff_dupl_max//diff_max;
             end;
   diff_dupl_min=T(diff_dupl_min);
   diff_dupl_max=T(diff_dupl_max);
   diff_strd=diff;
     do r=1 to m;
      do s=1 to n_b;
          if diff_dupl_max[r,s]=diff_dupl_min[r,s] then
             diff_strd[r,s]=0;
          else do;
                 diff_strd[r,s]=(diff[r,s]-diff_dupl_min[r,s])/
                      (diff_dupl_max[r,s]- diff_dupl_min[r,s]);
               end;
      end;
     end;
   dist=sqrt(diff_strd[##,]);
   D[i,]=dist;
end;
```

Nach der Bestimmung der Koeffizienten $d(a_i,b_j)$ kann das Problem *(LP)* mittels klassischer Methoden der Optimierung wie z. B. dem bekannten Simplex-Algorithmus oder spezieller auf Zuordnungsprobleme zugeschnittener Verfahren gelöst werden. Alternativ kann ein vorhandener LP-Solver wie etwa die in SAS/IML verfügbare Prozedur *linprog* eingesetzt werden. Bei großen Dateien (beispielsweise bei der Analyse amtlicher Steuerdaten) wird aus Effizienzgründen empfohlen, Näherungsverfahren zu implementieren.

Bei der entsprechenden SAS/IML-Implementierung kann zur Lösung des linearen Zuordungsproblems auf die Prozedur *linprog* zurückgegriffen werden. Hierzu ist zunächst eine Anpassung der Nebenbedingungen erforderlich:

```
obj=rowvec(D);     /* Koeffizienten der Zielfunktion */
coef=j(2#n,n#n,0);
     do k=1 to n;
        do i=1 to n;
           coef[k,i+(k-1)#n]=1;
        end;
     end;
     do k=n+1 to 2#n;
        do i=1 to n;
           coef[k,k-n+(i-1)#n]=1;
        end;
     end;
rel=j(2#n,1,'=');  /* Vergleichsoperatoren der Bedingungen */
rhs=j(2#n,1,1);    /* Rechte Seite der Gleichungen */
run linprog(names, obj, 'min', coef, rel, rhs, activity);
zuord=j(1,n,0);    /* Zeilenvektor der Zuordnungen */
     do i=1 to n;
          zuord[1,i]=activity[((i-1)#n)+i];
     end;
Anzahl_Treffer=sum(zuord);
```

Eine alternative Näherungslösung lässt sich durch die in empirischen Untersuchungen (siehe Lenz 2006) bewährte Näherungsheuristik GREEDY bestimmen, die hier zunächst im Pseudocode notiert wird. Die Idee besteht in einer sukzessiven Auswahl von Paaren kleinstmöglicher Distanz. Die Prozedur endet, wenn eine der beiden Datenquellen abgearbeitet ist.

```
Beginn GREEDY
     Sortiere die Distanzen aufsteigend in die Liste L
     Solange L nichtleer ist führe aus:
          Betrachte das erste Element d(a_i,b_j) von L und ordne a_i
          und b_j einander zu
          Entferne alle Elemente d(a_r,b_s)  mit r = i oder s = j
     aus L
Ende GREEDY
```

Solche-Heuristiken werden in der Praxis oftmals aufgrund ihrer einfachen Implementierung und des gegenüber dem Verfahren zur Bestimmung einer optimalen Lösung geringeren Rechenaufwandes vorgezogen. Im vorliegenden Fall müssen die in der Matrix *D* gespeicherten Distanzen für die Weiterverarbeitung zunächst in eine $(n_a+n_b)x2$ –Matrix

geschrieben werden. Der zugehörige überschaubare Auszug aus dem SAS/IML-Code lautet hier:

```
v=j(n_a#n_b,2,0);
do r=1 to n_a;
   do s=1 to n_b;
      v[D[r,s],1]=r;
      v[D[r,s],2]=s;
   end;
end;
```

5.3 Erste Ergebnisse von Matchingexperimenten für die Monatsberichte

Es wurden erste Matchingexperimente für die ca. 36.000 im Monatsbericht 2001 enthaltenen Einbetriebsunternehmen durchgeführt. Aus der kommerziellen Markus-Datenbank[4] haben wir Zusatzwissen für rund 9.000 dieser Einheiten. Als Blockvariablen wurden der Betriebssitz (alte/neue Bundesländer), der Wirtschaftszweig (Zweisteller) und die Anzahl der tätigen Personen zusammengefasst zu 6 Größenklassen verwendet. D. h.: Es werden nur die Unternehmen miteinander verglichen, die hinsichtlich dieser kategorialen Merkmale übereinstimmen. Die Überschneidungsmerkmale für die Zuordnung waren die Anzahl der tätigen Personen und der Umsatz. Bis jetzt wurde das Szenario nur für einen einzigen synthetischen Datensatz durchgeführt. Das Ergebnis: Nur vernachlässigbare 18 der etwa 9.000 Einheiten wurden richtig zugeordnet. Das sind 0,2 %.

Um zu prüfen, ob dieses Ergebnis vertrauenswürdig ist, haben wir den Anteil der übereinstimmenden Werte für die Blockvariablen in den originalen und den synthetischen Daten berechnet. Der zweistellige WZ-Code stimmt nur für 23,6 % der Einheiten in beiden Quellen überein, die Beschäftigtengrößenklasse in 80,2 % und der Sitz ist für 77,5 % der Betriebe identisch. Die Kombination aus allen drei Merkmalen bleibt für 14,7 % der Einheiten unverändert. Dies ist noch keine vollständige Erklärung, warum so wenige Betriebe korrekt zugeordnet wurden. Daher bedarf es weiterer Untersuchungen.
Ein wesentlicher Kritikpunkt an unserer bisherigen Vorgehensweise ist, dass ein potentieller Datenangreifer weiß, dass die kategorialen Merkmale teilweise verändert sind. Deshalb könnte er sich zumindest für eine Identifikation der größeren Einheiten (ab 1000 / 2000 Beschäftigten) nur auf die numerischen Variablen verlassen. Dieses Szenario muss deshalb auch noch betrachtet werden. Ferner könnten weitere Strategien denkbar sein, die speziell für synthetische Daten zielführend sind. In diesem noch relativ jungen Bereich ist noch ein großer Forschungsbedarf vorhanden.

[4] Die von Creditreform und Bureau van Dijk herausgegebene Markus Datenbank enthält umfangreiche Informationen zu ca. 1 Million deutscher und österreichischer Unternehmen. Eine Beschreibung des Produktes findet sich unter http://www.creditreform.de/Ressourcen/PDF/Produkte/6_Direktmarketing-Service/Business_Marketing/Broschuere_MARKUS_Datenbank.pdf

Literatur

[1] M. Brandt, M. Zwick (2009): infinitE – Eine informationelle Infrastruktur für das E-Science Age. Wirtschaft und Statistik 7/2009, 670-675.

[2] Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik (Hrsg.): Wege zu einer besseren informationellen Infrastruktur. Nomos Verlagsgesellschaft Baden-Baden, 2001.

[3] R. Lenz (2006): Measuring the Disclosure Protection of Micro Aggregated Business Microdata. An Analysis Taking as An Example the German Structure of Costs Survey. Journal of Official Statistics 22(4), 681-710.

[4] R. Lenz (2009): Défis méthodiques lors de la réalisation de l'accès aux données économiques allemandes par la téléinformatique automatisée. 41ème Journées de Statistique de la Société Francaise de Statistique (SFdS), Université Victor Segalen, Bordeaux.

[5] T. E. Raghunathan, J. M. Lepkowski, J. Van Hoewyk, P. Solenberger (2001): A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. Survey Methodology 27(1), 85-95.

[6] T. E. Raghunathan, J. Reiter, D. B. Rubin (2003): Multiple Imputation for Statistical Disclosure Control. Journal of Official Statistics 19(1), 1-16.

[7] J. P. Reiter (2003): Inference for partially synthetic, public use microdata sets. Survey Methodology 29, 181-188.

[8] D. B. Rubin (1993): Statistical Disclosure Limitation. Journal of Official Statistics 9(2), 461-468.