УДК 316

IMPUTATION OF MISSING DATA IN A NETWORK-PANEL WITH UP TO 8 WAVES

W. Sodeur

Wolfgang Sodeur, University Duisburg-Essen
Wolfgang.Sodeur@t-online.de

Статья поступила в редакцию 22.08.2019
Статья принята к публикации 28.10.2019

The University of Duisburg-Essen conducted a survey of student newcomers on their impressions of the first nine weeks of study. Respondents were asked questions about their socio-demographic characteristics, about their attitude to the educational activities and departments of the university, which they had had time to get acquainted with during the first weeks of study, as well as about the number of friendly contacts that new students had managed to establish by that time. The survey used a random sample using a name generator.

At some stages of data collection, it was found that information was insufficient, in particular, some students (about 200 persons) lacked information about the number of friendly contacts established during certain weeks. In such conditions, dynamic analysis of the development of social network relations becomes impossible.

In such situations, usually, imputation is applied to fill gaps by missing data. Here however, we use imputation methods to reproduce data which are already observed. This facilitates the possibility to analyse consequences of different imputation methods.

The method of imputation described here allows us to solve various problems. For this purpose, parts of the data obtained are first processed separately and then summed up using appropriate weights. The choice of different weights allows the procedure for compensating for missing data to be adapted to the specific purpose of the study. In order to make it easier to understand how the method of imputation works, we use only a small part of the data set. The imputations are then compared with the original/observed network data.

The data set and AWK programs are freely available at: Wolfgang.Sodeur@t-online.de

Comparative analysis results can be downloaded from www.uni-due.de/Sodeur/ together with the NN18IMP.zip file.

Ключевые слова: missing data, network panel, imputation, imputation tests.

Informationen zur Datensammlung

Studienanfänger wurden während der ersten 9 Wochen ihres Studiums an der Universität befragt. Sie gaben dabei zahlreiche Informationen über ihre Aktivitäten am Anfang des Studiums.

Während der ersten 5 Wochen des Studiums füllten sie an allen Werktagen die Formulare eines zum Teil standardisierten Tagebuches aus. Bedeutsam dabei waren vor allem ihre Kontakte zu anderen Studienanfängern. Kontakte zu denselben Studienanfängern aus unterschiedlichen Situationen wurden teilweise zusammengefasst und bilden für die Wochen 1-5 die Netze Nr. 18–25 (siehe das Daten-Archiv CNETZ von Lothar Krempel).

Zusätzlich wurden personale Kontakte der befragten Studierenden durch 15 «Namens-Generatoren» erfragt. Die erfragten Kontakte stehen überwiegend im Zusammenhang mit dem Studium und mit den besonderen Problemen am Beginn eines Studiums. Die entsprechenden Netz-Daten sind in CNETZ unter Nr. 1–15 dokumentiert und fallen je nach Generator in unterschiedlicher Verteilung in die Wochen 2–9.

Zwei weitere Namens-Generatoren stehen im Zusammenhang mit zum Teil studienbezogenen Büchern und sind in CNETZ unter Nr. 16–17 dokumentiert (vgl. H.J. Hummell und W. Sodeur 1984, Kapitel 2 unter www.uni-due.de/Sodeur/).

Alle sonstigen Daten, die nicht mit personalen Beziehungen zu bestimmten Personen verbundenen sind, werden in einem späteren Abschnitt kurz beschrieben.

Fehlende Daten über personale Beziehungen

Im Winter-Semester 1978/79 wurden dem Fachbereich Wirtschaftswissenschaft einer deutschen Universität durch eine zentrale Verteilungs-Organisation etwa 270 Studienanfänger zugewiesen.

Etwa 240 dieser Studierenden schrieben sich formal als Studienanfänger an der Universität ein oder waren zumindest während der ersten Wochen am Fachbereich anwesend.

208 von diesen 240 Studierenden nahmen an zumindest einer der 9 Wochen an der Befragung teil.

182 von diesen 208 Studierenden gaben in der 9. Woche ihre «persönliche Indentifikations-Liste» ab. Ohne diese Liste ist es unmöglich zu entscheiden, ob unterschiedliche Studierende dieselben oder unterschiedliche andere Studierende in ihren Fragebögen genannt haben. Die 26 Studierenden mit fehlenden Identifikations-Listen

müssen deshalb bei allen Untersuchungen der Beziehungsnetze ausgeschlossen werden.

Etwa 40 von den verbleibenden 182 Studierenden haben in mindestens einer Woche keine einzige andere Person in Tagebüchern oder nach Befragung mit Namens-Generatoren genannt. Der Grund dafür ist nicht klar: Manchmal mag es Abwesenheit z.B. durch Krankheit gewesen sein, manchmal auch bewusste Verweigerung von Angaben.

Wir betrachten alle diese Fälle gleichermaßen als «fehlende Daten». Ersatz durch Imputation ist hier besonders dringlich, wenn eine Analyse der dynamischen Entwicklung des Beziehungsnetzes angestrebt wird (z.B. mit Siena bzw. RSiena in der jeweils aktuellen Version, siehe Internet-Seite von Tom Snijders).

Für die fehlenden Angaben der 40 Studierenden wäre in den betreffenden Wochen die Imputation ihres jeweils ganzen persönlichen Beziehungsnetzes unter den Studienanfängern nötig.

Vielleicht fehlen zusätzlich auch die Beziehungsdaten einzelner Wochen für weitere etwa 30 Studenten aufgrund fehlerhafter Zuordnung ihrer Daten. Bei der Datenerhebung waren Daten auf der einen Seite und Namen sowie Adressen auf der anderen Seite streng voneinander getrennt. Namen sowie Adressen wurden von einer unabhängigen Institution (Treuhänder) verwaltet. Verbindungen zwischen Daten und Namen gab es nur über die Nummern der «Identifikations-Listen». Der Treuhänder musste die Identifikations-Nummern jedes der Studierenden über die 9 Wochen zusammenführen und auch mit den Nennungen von Beziehungs-Personen dieser Studierenden verknüpfen (vgl. H.J. Hummell und W. Sodeur 1984, S. 18ff). Mögliche Fehler bei der Verbindung dieser Identifikations-Nummern über Wochen führen zur falschen Zuordnung aller Daten des betreffenden Studierenden in dieser Woche, darunter auch aller «ausgehenden» Beziehungen.

Mögliche Fehler dieser Art wurden wie folgt identifiziert: Wir haben von jedem einzelnen Studenten die «ausgehenden» und die «eingehenden» Beziehungen zwischen den Wochen miteinander verglichen, und zwar in allen verfügbaren Netzen. In jenen 30 Fällen wich das Muster der «ausgehenden» Beziehungen in einer Woche sehr deutlich vom Muster in anderen Wochen ab. Gleichzeitig aber blieb das Muster der «eingehenden» Beziehungen (diese wären vom o.gen. Fehler nicht betroffen) weitgehend gleich. Daraus haben wir abgeleitet, dass die Daten des betreffenden Studierenden in dieser Woche vermutlich falsch zugeordnet wurden. Entsprechend werden die Daten des betreffenden Studierenden in dieser Woche als fehlend (missing data) erklärt. Für diese etwa 30 Studierenden gilt also ebenfalls ein besonders dringender Bedarf, die Lücken durch Imputation zu schließen.

Imputation fehlender Netzwerk-Daten zu einzelnen Wochen

Methoden zur Imputation fehlender Daten werden in der Literatur relativ häufig diskutiert. Oft steht dies im Zusammenhang mit der Analyse dynamischer Prozesse im Zeitverlauf. Fehlen dabei nur wenige Daten zu einzelnen Zeitpunkten Tx, so versucht man diese durch «optimierende Verfahren» zu ersetzen. Oft erfolgt dies durch (in weiterem Sinne) «multivariate Regressionsverfahren», bei denen einzelne fehlende Daten durch die optimale Kombination möglichst vieler vorhandener Daten ersetzt werden (siehe u.a. Bernhard Baltes-Götz 2013; Karen Steindorf und Oliver Kuß 2011).

Weiterführenden Fragen nach Möglichkeiten zur «multiplen» Imputation fehlender Werte, bei denen die Datenlücken nicht durch jeweils einen einzigen Wert, sondern durch dessen vermutete Verteilung ersetzt werden, spielen in beiden genannten Arbeiten eine große Rolle. Sie sind auch mit den hier behandelten Verfahren zu bearbeiten, werden aber in diesem Artikel nicht explizit verfolgt (vgl. Bernhard Baltes-Götz 2013; Karen Steindorf und Oliver Kuß 2011; s.u. Abschnitt «Ausblick»).

Fehlen dagegen zu bestimmten Zeitpunkten Tx alle Netzdaten einer Person, also Aussagen zu allen möglichen vorhandenen oder nicht vorhandenen «ausgehenden Beziehungen», so müssten im vorliegenden Fall gleichzeitig fehlende Daten über die mehr als 200 «potentiellen» Kontakte dieser Person erzeugt werden. Mit den üblichen Verfahren der Imputation ist dies nicht möglich.

Ersatzweise sucht man in solchen Fällen nach Personen mit vorhandenen Netzdaten, die zu anderen Zeitpunkten (T # Tx) möglichst große Ähnlichkeit zu der Person mit fehlenden Daten aufweisen («nearest neighbor»). Die vorhandenen Netzdaten dieser Ersatzperson dienen dann im Zeitpunkt Tx zur Imputation. Dies führt allerdings zu Problemen, wie frühere Untersuchungen gezeigt haben. Die «nächsten Nachbarn» sind oft zentrale Personen im Beziehungsnetz. Nimmt man sie häufig als Vorlage für Imputationen, so werden viele Struktureigenschaften des Netzes verzerrt, im vorliegenden Fall z.B. die Transitivität des Netzes erhöht (vgl. nicht veröffentlichte Poster auf der Sunbelt Conference im Juli 2018 in Utrecht und im September 2018 an der Lobachevsky Universität in Nizhny Novgorod: www.uni-due.de/Sodeur/).

Dieser verzerrende Effekt lässt sich vermeiden oder zumindest verringern, wenn man statt nur eines «nächsten Nachbarn» mehrere relativ nahe Nachbarn als Vorlage für die Imputation benutzt.

Die hier gewählten Verfahren zur Imputation von Netzdaten wählen deshalb als Daten-Grundlage stets die drei relativ nächsten Nachbarn jedes Studierenden. Die ausgehenden Beziehungen dieser drei nächsten Nachbarn werden kombiniert. Daraus wird eine Zufallsauswahl mit einem Außengrad entsprechend dem arithmetischen Mittelwertes der 3 zugrunde liegenden Außengrade gezogen. Das Ergebnis dieser Änderung des Verfahrens ist, dass die durch Imputation gewonnenen Netzwerke in ihrer Transitivität in geringerem Umfang von den ursprünglich beobachteten Netzwerken abweichen. Eine systematische Prüfung einer unterschiedlichen Zahl relativ nächster Nachbarn (von 1–10) ergab eine besonders gute Lösung mit 3 relativ nahen Nachbarn (vgl. A. Znidarsic u.a. 2018).

Datenbasis zur Ermittlung der jeweils «nächsten Nachbarn»

Wie schon in der Einleitung kurz beschrieben wurde, liegen aus dem genannten Projekt der «Studienanfänger» sehr umfangreiche Daten über die beteiligten Studenten vor. Diese Daten werden hier ausführlicher beschrieben. Sie werden nach inhaltlichen Bereichen gegliedert und jeweils paarweise – d.h. für jeden der 182 befragten Studierenden im Vergleich zu den jeweils anderen 181 Studierenden – zu einem von 0-1 normierten Distanzmaß zusammengefasst.

Dies führt zu einer übersichtlichen Darstellung der vorhandenen Datenbereiche. Die paarweisen Distanzen zwischen Studierenden können anschließend mit Gewichten für jeden der Bereiche versehen und dann zusammengefasst werden. Damit wird die Möglichkeit geschaffen, die Suche nach «nächsten Nachbarn» und damit die Steuerung der anschließenden Imputation verschiedenen inhaltlichen Zielsetzungen anzupassen.

In diesem Artikel nutzen wir die Möglichkeit zu beliebigen Gewichtungen nur beispielhaft: Zur Demonstration werden aus den verfügbaren Datenbereichen drei Gruppen ausgewählt. Anschließend wird gezeigt, inwieweit die auf dieser Basis erzeugten imputierten Beziehungsnetze mit den ursprünglich erhobenen Netzen übereinstimmen. Die weitere, durch bestimmte Kombinationen von Gewichten gesteuerte Imputation bleibt interessierten Lesern vorbehalten. Die entsprechenden Programme dazu sind frei verfügbar. Zu verändern sind jeweils nur die Gewichte im Einleitungs-Kapitel («BEGIN») des Programms Crall2.awk.

Zunächst aber werden alle verfügbaren – also nicht nur die in diesem Aufsatz verwerteten – Datenbereiche und die daraus entstehenden Dateien mit paarweisen Distanzen kurz beschrieben.

Wahl der paarweisen Beziehungen zwischen Studierenden zur Suche nach nächsten Nachbarn;

Beziehungsnetze und daraus abgeleitete Distanz-Listen Nr. 1-4

Die 116 in CNETZ enthaltenen Beziehungsnetze der 182 Studierenden beruhen zum Teil auf Tagebuch-Aufzeichnungen (Netz-Nummern 18-25) und zum Teil auf zahlreichen «Namens-Generatoren» (Netz-Nummern 1–17), mit denen sowohl studienbezogene wie auch persönliche Kontakte der Studierenden erhoben wurden. Diese Daten wurden über alle verfügbaren Wochen zu nur 4 Netzen zusammengefasst.

Auf der Grundlage der Namens-Generatoren und der Tagebuch-Aufzeichnungen gibt es jeweils ein Netz mit allen «ausgehenden» und jeweils ein Netz mit allen «eingehenden Wahlen». Als Wahlen bezeichnen wir hier persönliche Nennungen eines Studierenden über andere Studierende.

Ohne Zwischenschritte zu nennen, beschreiben wir kurz die Entwicklung der 4 genannten Distanz-Dateien. Als Beispiel werden dabei nur die Nennungen der ersten 3 Studierenden angeführt. Es handelt sich dabei um die 2. Woche, den Namens-Generator 1 und das 9. Netzwerk (von insgesamt 116 aus CNETZ):

«Innerhalb unseres Fachbereiches gibt es verschiedene Moeglichkeiten der Faecherwahl und innerhalb dieser Faecher wieder verschiedene Moeglichkeiten, das Studium individuell auszurichten. Uns interessiert nicht, was Sie hierueber wissen, sondern, ob Sie darueber etwas von Mitstudenten in der vergangenen Woche gehoert haben».

```
# ka11.txt
Week/Net 201 NetSequ.  9 Length  718
  1 340
  2  15
  3  19  20 117 153 302
usw...
```

Anstelle der Beziehungs-Matrix mit der Dimension 182*416 wird hier eine «Listen-Darstellung» des Beziehungsnetzes gewählt. Die Studierenden 1, 2, 3 (entsprechend den ersten 3 Zeilen) haben die anschließend aufgeführten Studierenden zum Namens-Generator 1 genannt. Alle nicht-genannten Personen (mit «0» in der Matrix) kommen in der Listen-Darstellung nicht vor. Die Identifikationsnummern 340 (Zeile 1) und 302 (Zeile 3) verweisen nicht auf Studienanfänger, sondern auf andere Personen aus dem Umfeld der Studierenden. Diese Personen werden nicht persönlich, sondern nur «kategorial» erfasst (wie z.B. «Freund aus früherer Schulklasse», «Mutter», «Bruder»). Solche Nennungen werden bei der Suche nach nächsten Nachbarn nicht mehr berücksichtigt.

Sobald die Nennungen zu allen Namens-Generatoren (1–17) und allen Wochen (maximal 8 Wochen, insgesamt 76 von 116 Netzen aus CNETZ) zusammengefasst sind, ergibt sich die folgende Datei:

```
# ka11b.txt
#Ka ID NV Nennungen...
11   1 411 13:1 158:21 196:1 200:16 340:8 341:3 342:3 358:6
11   2 411 15:28 65:5 106:16 181:3 267:23 302:19 340:11 341:1 358:3 383:1
11   3 411 19:6 20:8 117:6 153:24 168:3 174:5 302:10 394:1
usw...
```

In der Datei ka11b.txt sind wieder als Beispiel die kumulierten Nennungen der drei ersten Befragten zusammengefasst. Nach der Quelle der Daten (Kartenart 11) folgen die Nummern der Befragten (hier nur 1,2,3 von 1-182), und die Zahl der insgesamt vorkommenden persönlichen und kategorialen Nennungen (maximal 411). Es handelt sich also um eine Matrix mit 182 Zeilen (Befragten) und 411 Spalten (genannte Studienanfänger bzw. Personen-Kategorien), von denen später aber nur die ersten 272 (auf Personen bezogenen) Spalten berücksichtigt werden.

Nach der Spaltenzahl (411) folgen vor den Doppelpunkten die Nummern der genannten Beziehungs-Personen und danach die Häufigkeiten, mit denen diese Personen auf alle 17 Namens-Generatoren und in allen Wochen genannt wurden.

13:1 in der ersten Zeile bedeutet also, dass Person 13 von Person 1 in allen 76 Netz-Daten insgesamt nur einmal genannt wurde, 158:21 entsprechend, dass Person 158 insgesamt 21 mal genannt wurde.

Vergleicht man nun diese kumulierten Nennungen von Beziehungs-Personen im Bereich der Nummern 1-272 (die nur kategorialen Nennungen werden ausgeschlossen) paarweise zwischen allen 182 befragten Studierenden, so ergibt sich die folgende Datei ka11co.txt. Am Rande sei erwähnt, dass bei der Inversion der Beziehungs-Matrix, wenn also Zeilen und Spalten vertauscht werden, in jeder Zeile (nun von 1-272) alle bei der zuerst genannten Person «eingehenden» Nennungen genannt werden. So entsteht dann die inverse Matrix ka11cin.txt.

```
###Ka11co
# max.no.of targets (<=maxid): 22 272
# ka,max1(sources),max2(targets): 11 182 272
#ka i  j sumd n1 n2 n12 npair diff(x)    diff(xn)
11  1  1    0  0  0  4 0    0.00000    0.00000
11  1  2  114  4  5  0 0   12.66667    1.00000
11  1  3   91  4  6  0 0    9.10000    1.00000
usw...
```

In den Zeilen der Überschrift werden die Quelle der Daten (wieder Kartenart 11), die maximale Zahl unterschiedlicher Nennungen (22 von 272 möglichen Nennungen), und die Dimension der Matrix (182 Zeilen, 272 Spalten) genannt. Schließlich werden auch die einzelnen Spalten der Tabelle bezeichnet, nämlich

– die Kartenart (ka),

– die ID-Nummern von jeweils zwei befragten Personen (i,j),

– die Summe (sumd) der zwischen beiden insgesamt bestehenden Häufigkeits-Unterschiede (bzw. ihrer „Distanz“),

– die Zahl der nur von der ersten Person genannten (von i, n1) bzw.

–nur von der zweiten Person genannten (von j, n2) bzw.

– von beiden gleichzeitig genannten Beziehungs-Personen (von i und j, n12), und

– schließlich die pro Nennung beider Personen durchschnittlich entstandene Differenz der Häufigkeiten bzw. ihre Distanz (diff(x): sumd/(n1+n2+ 2*n12)).

Auf die beiden übrigen Spalten (npair und diff(xn) wird hier nicht eingegangen.

Zwischen den Befragten 1 und 2 ergibt sich also eine Häufigkeits-Differenz (sumd) von 114 bei insgesamt 4+5=9 Nennungen, daraus wird eine mittlere Differenz von 114/(4+5)=12.66667.

Die relativen Distanzen eines der 182 Befragten zu allen anderen 181 Befragten sind die Basis zur Suche nach seinen nächsten Nachbarn. Diese nächsten Nachbarn liefern im Fall fehlender Daten die Quelle für Ersatzdaten, wenn z.B. von einem bestimmten Befragten zu einem bestimmten Namens-Generator und einer bestimmten Woche keine Daten vorliegen.

Die oben genannten relativen Distanzen in der Spalte diff(x) könnten diese Suche leiten, solange sie allein für die Suche nach nächsten Nachbarn herangezogen werden sollen.

Sobald aber mehrere solcher Bereiche zusammengefasst werden, muss über ihre relative Gewichtung nachgedacht werden. Schon allein auf Grund der direkten Nennungen der Befragten in Tagebüchern und auf die 17 Namens-Generatoren hatten wir – trotz Zusammenfassung – insgesamt 4 solcher Distanz-Listen genannt. Zur Zusammenfassung der Distanz-Listen aus mehreren Bereichen müssen diese also nochmals standardisiert werden. Die Distanz jedes einzelnen Befragten zu allen übrigen 181 Studierenden muss über verschiedene Bereiche hinweg vergleichbar sein.

Diese Standardisierung wird getrennt für die paarweisen Distanzen jedes Befragten einzeln durchgeführt. Von den 181 mittleren Distanzen der Datei k11co.txt (in Spalte diff(x)) eines Befragten wird deshalb der jeweils kleinste und größte Wert gesucht. Beim Befragten mit i=1 sind dies die Distanz-Werte 3.11111 (gegenüber j=84) und 24.625 (gegenüber j=42).

Zur Standardisierung werden nun diese 181 Distanz-Werte des Befragten i=1 so transformiert, das sein nächster Nachbar (mit kleinster Distanz, hier j=84) den Wert 0.0 erhält und sein entferntester Nachbar (mit größter Distanz, hier j=42) den Wert 1.0.

Die standardisierte Distanz des ersten Befragten (i=1) zum Zweiten (j=2) und Dritten (j=3) errechnet sich also aus den Distanzen diff(x) der Datei k11co.txt (s.o.) wie folgt:

Norm-Distanz(1->2)=(12.66667-3.11111)/
(24.625-3.11111)= 0.44416
Norm-Distanz(1->3)=(9.10000-3.11111)/
(24.625-3.11111)= 0.27837

Entsprechend werden die weiteren Distanzen des ersten Befragten (i=1) zu allen anderen Studierenden berechnet. Gleiches geschieht mit den jeweils 181 Distanzen aller anderen Befragten. Die Ergebnisse stehen in Datei ka11coS.txt.

Alle weiteren Distanzen, die auf direkten Nennungen der befragten Studierenden beruhen, werden auf gleiche Weise zu standardisierten Distanz-Listen umgeformt. Es entstehen damit die folgenden 4 standardisierten Distanz-Listen:

Kumulierte Nennungen zu allen Namens-Generatoren und allen Wochen
(Quelle: Ka11.txt, 76 Netze, maximal 8 und durchschnittlich etwa 6 Wochen pro Netzart)
Ka11coS.txt (outgoing links) Distanzen aus Matrix 182*272
Ka11cinS.txt (incoming links) Distanzen aus invertierter Matrix 272*182
Kumulierte Nennungen aus allen Tagebuch-Aufzeichnungen und allen Wochen
(Quelle: ka41.txt, 40 Netze, jeweils 5 Wochen pro Tagebuch-Kategorie)
Ka41coS.txt (outgoing links) Distanzen aus Matrix 182*272
Ka41cinS.txt (dito incoming) Distanzen aus invertierter Matrix 272*182

Wahl des Positionen-Zensus aus allen 116 Netzen
zur Suche nach nächsten Nachbarn
und daraus abgeleitete Distanz-Liste Nr. 5

Bei der Suche nach nächsten Nachbarn werden nun nicht nur die Muster ausgehender direkter Beziehungen der befragten Studenten berücksichtigt, sondern auch alle möglichen Kombinationen dieser Beziehungen in ihrer «triadischen Umgebung».

Wir wählen dazu nicht den Triaden-Zensus, wie dies häufig und meist fehlerhaft in der Literatur beschrieben wird. Der Grund dafür ist, dass einige der Triaden-Typen die «triadische Umgebung» der jeweils drei Personen sehr unterschiedlich kennzeichnen. Dazu sei als extremes Beispiel die Triade 030T von insgesamt 16 Typen des Triaden-Zensus angeführt (vgl u.a. Hans J. Hummell und Wolfgang Sodeur 1987, S.134):

Triade 030T 2<--1-->3-->2

Die drei dazu gehörenden Personen (1,2,3) haben

•zwei ausgehende und keine eingehende Beziehung (1),
•zwei eingehende und keine ausgehende Beziehung (2),
•eine ausgehende und eine eingehende Beziehung (3).

Derselbe Triaden-Typ beschreibt also keineswegs gleiche, sondern denkbar unterschiedliche triadische Umgebungen der drei Personen und ist deshalb ungeeignet zur eindeutigen Kennzeichnung individueller Umgebungen.

Statt des Triaden Zensus wählen wir deshalb den Positionen-Zensus, der eine Häufigkeitsverteilung über alle 36 möglichen Positionen eines Studierenden in der jeweiligen triadische Umgebung vornimmt (vgl. H.J. Hummell und W. Sodeur 1987, S.188). Ron Burt (1990) hat später dieselben 36 Typen auf besser lesbare und verständlichere Weise beschrieben, sie allerdings insgesamt «role census» genannt und anders numeriert:

Numerierung der Positionen: Hummell/Sodeur(1) --> Burt (2)

---

(1) 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
(2) 1 2 3 4 6 8 9 21 22 31 23 24 26 28 29 5 10 7 32

(1) 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36
(2) 34 33 35 25 30 36 27 11 12 13 14 16 18 19 15 20 17

Der Positionen-Zensus zählt die Häufigkeiten aller möglichen 36 triadischen Umgebungen von jedem der 182 Studierenden. Ihre Summe np in einem Netz von n Personen beträgt für jede der n Personen np = (n-1)*(n-2)/2 und entspricht damit der Zahl triadischer Umgebungs-Typen, die von den jeweils anderen (n-1) bzw. (n-

2) Personen gebildet werden. Auf der Suche nach dem einen «nächsten Nachbarn» jedes Studierenden bzw. den (hier) 3 «relativ nächsten Nachbarn» muss der Positionen-Zensus jedes der n=182 Studierenden mit dem entsprechenden Zensus der anderen n-1=181 Studierenden verglichen werden. Bei diesem Vergleich bleibt jedoch die erste Position des Zensus unberücksicht. Sie ist «leer» in dem Sinne, dass keine einzige Beziehung zwischen den 3 Studierenden besteht, kommt aber bei den hier betrachteten, meist «dünnen» Beziehungsnetzen besonders häufig vor.

Ihre Häufigkeit darf auch deshalb nicht genutzt werden, weil sie (infolge der konstanten Summe aller Positions-Häufigkeiten) linear abhängig von den übrigen 35 Positions-Häufigkeiten ist.

Vor jedem Vergleich zweier Studierender werden die Positionen-Zensus aller 116 Netze (aus CNETZ) zusammengefasst. Die Summe der Differenzen unter den verbleibenden 35 Positionen wird dann durch die Zahl der besetzten (Häufigkeit >0) Positionen geteilt. Damit werden nur jene Positionen berücksichtigt, die bei mindestens einem der beiden Studierenden mindestens einmal vorkommen.

Paarweise, auf Werte 0.0-1.0 standardisierte Distanzen jedes der 182 Studierenden mit den jeweils übrigen 181 anderen sind in der Datei Ka1141po.txt gespeichert.

Wahl sonstiger Datenbereiche, die nicht auf Beziehungsnetze bezogen sind, zur Suche nach «nächsten Nachbarn» und daraus abgeleitete Distanz-Listen Nr. 6-37

Hier werden kurz alle übrigen, d.h. nicht auf Beziehungsnetze bezogenen Daten aus den angewandten Fragebögen beschrieben. Das sind einige wenige sozio-demographische Variablen, einige Einstellungen, die besuchten Lehrveranstaltungen für das 1. Semester, und die Einrichtungen der Universität, die während der ersten Wochen des Studiums besucht wurden.

Alle diese Daten wurden zum paarweisen Vergleich der Studierenden zunächst für jeden Bereich und jede Woche einzeln durch Korrelations-Koeffizienten zusammengefasst. Bei binären (0/1) Variablen diente dazu der Korrelations-Koeffizienten PHI, bei Rangordnungs-Daten der Rang-Korrelations-Koeffizient nach Spearman.

In beiden Fällen entstehen für jeden einzelnen der N=182 befragten Studierenden

N-1 =181 Koeffizienten, die seine relative Ähnlichkeit mit den jeweils anderen 181

Studierenden beschreiben. Diese Koeffizienten werden zur „Linearisierung“

quadriert und das Vorzeichen in umgekehrter Fassung hinzugefügt, wodurch aus

dem Ähnlichkeits- ein Distanz-Maß wird.

Schließlich werden die paarweisen Distanzen jedes Studierenden zu den

jeweils 181 anderen Studierenden wieder auf die Werte zwischen 0 (minimale Distanz)

bis 1 (maximale Distanz) normiert.

Distanz-Listen Nr. 6-21 mit Kennzeichen «a1» bzw. «a2»: In Woche 2 bis Woche 9 (bzw. an den jeweiligen End-Tagen der 8 Wochen 10,15,20,25,30,35,40,45) wurden die Studierenden gefragt, (a1) welche von 13 Institutionen der Universität sie besucht und (a2) welche von ebenfalls 13 «studien-bezogenen»Tätigkeiten sie ausgeführt hatten.

Wie oben beschrieben wurden zu beiden Bereichen pro Woche die paarweisen Ähnlichkeiten der Studierenden durch Phi-Koeffizienten ermittelt, danach zu Distanzen umgeformt und schließlich für jeden der 182 Studierenden in jeweils 181 normierte Distanzen im Bereich 0 («identisch») bis 1 («unter den 181 anderen Studierenden maximal ungleich») transformiert.

Für beide Bereiche und 8 Wochen liegen also 2*8=16 vollständige Sätze paarweiser

Distanzen vor.

Distanz-Listen Nr. 22-24 mit Kennzeichen «b»: In den Wochen 2, 4 und 7 (bzw. an End-Tagen 10, 20, 35) wurde von den Studierenden die Rangordnung (1: relativ größte Bedeutung; 18: relativ geringste Bedeutung) von 18 vorgegebenen Zielen für ihr Studium erfragt. Für jede dieser 3 Wochen wurden die paarweisen Ähnlichkeiten der Studierenden durch den Rang-Korrelations-Koeffizienten nach Spearman ermittelt und nach den genannten Regeln zu normierten Distanzen umgeformt. Für diesen Bereich liegen also 3 vollständige Sätze paarweiser Distanzen vor.

Distanz-Listen Nr. 25-27 mit Kennzeichen «c»: In den Wochen 3, 5 und 8 (bzw. an End-Tagen 15,25,40) wurde von den Studierenden die Rangordnung (1: relativ größte Bedeutung; 18: relativ geringste Bedeutung) von 18 vorgegebenen Kriterien für soziale Kontakte erfragt. Für jede dieser 3 Wochen wurden die paarweisen Ähnlichkeiten der Studierenden durch den Rang-Korrelations-Koeffizienten nach Spearman ermittelt und nach den genannten Regeln zu normierten Distanzen umgeformt. Für diesen Bereich liegen also 3 vollständige Sätze paarweiser Distanzen vor.

Distanz-Listen Nr. 28-31 mit Kennzeichen «d»: In den Wochen 2, 3, 4 und 5 (bzw. an End-Tagen 10,15,20,25) wurde gefragt, welche Informationsquellen (aus einer vorgegebenen Liste von 16 möglichen Quellen) die Studierenden für die

Wahl ihres Studienfaches herangezogen hatten. Für jede der 4 Wochen wurden die paarweisen Ähnlichkeiten der Studierenden durch Phi-Koeffizienten ermittelt und nach den genannten Regeln zu normierten Distanzen umgeformt. Insgesamt liegen für diesen Bereich also 4 vollständige Sätze paarweiser Distanzen vor.

Distanz-Listen Nr. 32-35 mit Kennzeichen «e»: In den Wochen 2, 3, 4 und 5 (bzw. an End-Tagen 10,15,20,25) wurde auch gefragt, welche Informationen (aus einer vorgegebenen Liste von 16 möglichen Informationen) die Studierenden für die Planung ihres «Stundenplans», das heißt der zu besuchenden Lehrveranstaltungen, herangezogen hatten. Für jede der 4 Wochen wurden die paarweisen Ähnlichkeiten der Studierenden durch Phi-Koeffizienten ermittelt und nach den genannten Regeln zu normierten Distanzen umgeformt. Insgesamt liegen für diesen Bereich also 4 vollständige Sätze paarweiser Distanzen vor.

Distanz-Liste Nr. 36 mit Kennzeichen «f»: In Woche 6 (bzw. am End-Tag 30) wurde gefragt, welche Lehrveranstaltungen (aus einer vorgegebenen Liste von allen 19 zur Wahl stehenden Lehrveranstaltungen/Dozenten) die Studierenden in den ersten Wochen des Studiums besucht hatten. Aus den Mustern der besuchten Veranstaltungen wurden die paarweisen Ähnlichkeiten der Studierenden durch Phi-Koeffizienten ermittelt und nach den genannten Regeln zu normierten Distanzen umgeformt. Dieser Bereich ist also durch einen vollständigen Satz paarweiser Distanzen vertreten.

Distanz-Liste Nr. 37 mit Kennzeichen «g»: In Woche 6 (bzw. am End-Tag 30) wurden auch 5 sozio-demographische Merkmale mit insgesamt 16 vorgegebenen Werte-Gruppen erhoben. Nach Transformation dieser Daten zu 16 Binär-Variablen wurden paarweise die Phi-Koeffizienten berechnet und nach den genannten Regeln zu normierten Distanzen umgeformt. Dieser Bereich ist also ebenfalls durch einen vollständigen Satz paarweiser Distanzen vertreten.

Insgesamt liegen also 5 Distanz-Listen auf der Basis von Daten über Beziehungsnetze und 32 Distanz-Listen auf der Basis anderer Daten vor. Alle 37 Distanz-Listen sind als Dateien auf der schon oben genannten Web-Seite www.uni-due.de/Sodeur/in der Sammel-Datei NN18IMP.zip zu finden.

Tabelle 1 enthält die Namen der Dateien aller 37 Distanz-Listen und Gewichte, die diesen Listen in verschiedenen Zusammenhängen zugewiesen wurden. Die Namen der 5 Dateien auf Basis von Netz-Daten wurden bereits oben genannt. Die Namen der 32 Dateien auf anderen Grundlagen sollen am Beispiel der zuletzt genannten Distanz-Liste aufgrund der erhobenen sozio-demographischen Merkmale erläutert werden. Sie trägt den Namen «CR930gA.txt».

Sie beginnt wie alle anderen Distanz-Dateien aufgrund der sonstigen Daten mit den Buchstaben «CR» und einer Ziffer «8» oder «9», die auf die Quelle der Daten verweist. #Fussnote: Es handelt sich um die «Kartenart» 8 oder 9. Die Datenerhebung fand zu einer Zeit statt, als noch «Lochkarten» als Datenträger dienten. Die ursprünglichen Daten und ihre Dokumentation sind ebenfalls über die genannte Webseite des Autors verfügbar. Dort sind auch alle Einzelheiten über die zur Erhebung genutzten Fragen dokumentiert. # In diesem Fall («CR9») folgt der End-Tag der betreffenden Erhebungswoche (hier für Woche 6 der End-Tag 30) und das für die sozio-demographischen Daten gewählte Kennzeichen «g». Das abschließende «A» des Datei-Namens (vor dem Datei-Typ «.txt») ist allen auf Werte 0–1 normierten Distanz-Listen der befragten Studierenden gemeinsam. Eine ebenfalls frei verfügbare Datei ohne dieses «A» enthält die noch nicht (pro befragtem Student) normierten Distanzen. Sie trägt in diesem Fall also den Namen «CR930g.txt».

Nun zu den ebenfalls in Tabelle 1 genannten «Gewichten» der einzelnen Bereiche. Bis zu der Spalte «all» enthält sie einen Auszug aus dem Kapitel «BEGIN» des Programms Crall2.awk, das bei Änderungen der Gewichte entsprechend gestaltet werden müsste. Die Spalte «all» enthält einen ersten, eher formal als inhaltlich begründeten Vorschlag zur Zusammenfassung aller Daten. Dabei dient als Gewicht die Zahl der Wochen, in denen die entsprechenden Befragungs-Instrumente eingesetzt wurden. Die folgenden Spalten enthalten alternative Vorschläge zur Auswahl von Teilmengen der verfügbaren Datenbereiche. Besonders einschneidende Beschränkungen der genutzten Datenbereiche sind in den Spalten mit Überschriften «nn, out, pos» gewählt. Diese werden als Beispiele mit ihren Konsequenzen für die Imputation von ganzen Netzwerken im folgenden Kapitel dargestellt.

In Spalte «nn» (no networks) werden die Gewichte aller Bereiche auf «0» gesetzt, die auf Netzdaten beruhen (Bereiche 1–5). Alle 32 sonstigen Datenbereiche werden dagegen «mit gleichen Gewichten» (=1) berücksichtigt.

In Spalte «out» (outgoing links) werden für die Suche nach nächsten Nachbarn nur die Netze mit ausgehenden Beziehungen auf Basis der Namens-Generatoren (Ka11coS.txt, Gewicht=6) und der

Tagebücher (Ka41coS.txt, Gewicht=5) entsprechend der Zahl der Wochen berücksichtigt, in denen diese Daten erhoben wurden.

In Spalte «pos» (position census) wird nur der kombinierte Positionen-Zensus der 116 Netze aus CNETZ (Ka1141po.txt) berücksichtigt. Im Grunde liegt hier dieselbe Datenbasis zugrunde wie in Spalte «out». Während dort aber nur die einfachen Muster der ausgehenden Beziehungen paarweise zwischen Studierenden verglichen werden, sind es hier (pos) die «triadischen Umgebungen» der Studierenden in Form des Positionen-Zensus.

Tabelle 1: Namen und alternative Gewichte der 37 Distanz-Listen
(Auszug aus Kapilel BEGIN des Programms Crall2.awk)

```
# The search for nearest neighbours uses the following variable groups:
# all: all available data; i/o: outgoing+ingoing links; nn: no network data;
# out: outgoing links only; pos: position census; -pos: all but position census

BEGIN {alli/o  nn out pos -pos(NN)
------------------------------------------------------------
#   links by "sociometric choices"
 filenam["ka11cinS.txt"]=6  =6  =0  =0  =0  =6
 filenam["ka11coS.txt"]= 6   =6  =0  =6  =0  =6
#   links by "diaries"
 filenam["ka41cinS.txt"]=5  =5  =0  =0  =0  =5
 filenam["ka41coS.txt"]=5  =5  =0  =5  =0  =5
#   position census, 116 networks
 filenam["ka1141po.txt"]=9   =0  =0  =0  =9  =0
#   university services visited
 filenam["cr810a1A.txt"]=1   =0  =1  =0  =0  =1
 filenam["cr815a1A.txt"]=1   =0  =1  =0  =0  =1
 filenam["cr820a1A.txt"]=1  =0  =1  =0  =0  =1
 filenam["cr825a1A.txt"]=1 =0  =1  =0  =0  =1
 filenam["cr830a1A.txt"]=1  =0  =1  =0  =0  =1
 filenam["cr835a1A.txt"]=1 =0  =1  =0  =0  =1
 filenam["cr840a1A.txt"]=1  =0  =1  =0  =0  =1
 filenam["cr845a1A.txt"]=1 =0  =1  =0  =0  =1
#   study-related activities
 filenam["cr810a2A.txt"]=1  =0  =1  =0  =0  =1
 filenam["cr815a2A.txt"]=1  =0  =1  =0  =0  =1
 filenam["cr820a2A.txt"]=1 =0  =1  =0  =0  =1
 filenam["cr825a2A.txt"]=1 =0  =1  =0  =0  =1
 filenam["cr830a2A.txt"]=1 =0  =1  =0  =0  =1
 filenam["cr835a2A.txt"]=1 =0  =1  =0  =0  =1
 filenam["cr840a2A.txt"]=1  =0  =1  =0  =0  =1
 filenam["cr845a2A.txt"]=1 =0  =1  =0  =0  =1
#   rank of goals for study
 filenam["cr910bA.txt"]=1 =0  =1  =0  =0  =1
 filenam["cr920bA.txt"]=1 =0  =1  =0  =0  =1
 filenam["cr935bA.txt"]=1 =0  =1  =0  =0  =1
#   rank of criteria for social relations
 filenam["cr915cA.txt"]=1 =0  =1  =0  =0  =1
 filenam["cr925cA.txt"]=1  =0  =1  =0  =0  =1
 filenam["cr940cA.txt"]=1 =0  =1  =0  =0  =1
#   info-sources, subjects for study
 filenam["cr910dA.txt"]=1  =0  =1  =0  =0  =1
 filenam["cr915dA.txt"]=1  =0  =1  =0  =0  =1
 filenam["cr920dA.txt"]=1  =0  =1  =0  =0  =1
 filenam["cr925dA.txt"]=1  =0  =1  =0  =0  =1
#   info-sources, planning time table
 filenam["cr910eA.txt"]=1  =0  =1  =0  =0  =1
 filenam["cr915eA.txt"]=1 =0  =1  =0  =0  =1
 filenam["cr920eA.txt"]=1 =0  =1  =0  =0  =1
 filenam["cr925eA.txt"]=1   =0  =1  =0  =0  =1
#   lectures visited (subject, lecturer)
 filenam["cr930fA.txt"]=1  =0  =1  =0  =0  =1
# 5 demographic data (transf.16 binary v.)
 filenam["cr930gA.txt"]=1 =0  =1  =0  =0  =1
...  }
------------------------------------------------------------
```

Drei Beispiele zur Imputation ganzer Beziehungsnetze
auf Basis einfacher Untermengen der verfügbaren Daten-Bereiche

Wie im vorausgehenden Abschnitt angekündigt, werden verschiedene Imputationen ganzer Netze als vereinfachende Beispiele unter Auswahl jener Distanz-Listen vollzogen, die sich auf folgende Daten beziehen: (1: nn) auf alle sonstige Daten (d.h. Netzdaten sind ausgeschlossen); vgl. Distanz-Listen 6-37 (2: out) auf alle Beziehungsnetze mit ausgehenden Wahlen; vgl. Ka11coS.txt (outgoing links): Namens-Generatoren und Ka41coS.txt (outgoing links): Tagebuch-Aufzeichnungen (3: pos) auf alle Positionen-Zensus auf Basis ausgehender Wahlen. vgl. Distanz-Liste Ka1141po.txt

Alle Auswertungen beziehen sich nur auf eines der 116 Beziehungsnetze aus CNETZ, nämlich das Netz Nr.12. Es handelt sich dabei um den Namens-Generator Nr.4, der in der zweiten Befragungswoche mit dem End-Tag 10 zum ersten Mal eingesetzt wurde:

«Wenn Sie eine Hausarbeit als Gruppenarbeit anfertigen, mit welchen Kommilitonen/Innen (einmal ganz abgesehen von den speziellen Fachkenntnissen) wuerden Sie gerne zusammenarbeiten?»

Die vier Tabellen beruhen auf Auswertungen mit dem (ebenfalls frei verfügbaren) Programm NETZDIAU.exe. Der Text ist zu Vergleichs-Zwecken wie in den Vorlagen auf Englisch geblieben.

Text und Tabellen je nach Umbruch beliebig anzuordnen

Die Eigenschaften der beobachteten Beziehungsnetze werden durch die nach drei Verfahren imputierten Netze sehr unterschiedlich reproduziert. Keines der Verfahren zur Suche nach nächsten Nachbarn bewährt sich dabei in allen Fällen gleich gut.

Der Positionen-Zensus bringt die Häufigkeit und die Verteilung der ausgehenden Wahlen den Ergebnissen der Befragten am relativ nächsten (Tabelle2). Das gilt auch für die Häufigkeits-Verteilung der eingehenden Wahl (Tabelle 3), obwohl hier die Abweichungen schon größer sind.

Die Anpassung der Häufigkeit und Verteilung der symmetrischen (gegenseitigen) Wahlen (Tabelle 4) gelingt dagegen mit allen drei Verfahren nur schlecht, der Positionen-Zensus bringt dabei sogar – neben den Netzdaten über ausgehende Wahlen – besonders schlechte Ergebnisse.

Gerade die Netze über ausgehende Wahlen bringen jedoch die besten Ergebnisse, wenn es um die Anpassung der Triplett-Transitivität geht (Tabelle 5). Hier hätte man eine bessere Anpassung durch den Positionen-Zensus erwarten können, der als einziger der hier genutzten Bereiche die Struktur der Beziehungen im Umfeld der Befragten berücksichtigt. Erklärt werden kann das enttäuschende Ergebnis des Positionen-Zensus vielleicht aus dem Umstand, dass die Triplett-Transitivität nur auf die Häufigkeiten von 2 der 8 Triplett-Typen zurückgeht. Auch die große Mehrheit aller berücksichtigten Positionen (triadischen Umgebungen), welche die Suche nach nächsten Nachbarn gesteuert haben, berücksichtigen nicht die Unterscheidung zwischen transitiven und intransitiven Strukturen.

Zusammenfassend kann man also feststellen, dass es für die Suche nach nächsten Nachbarn, die für die Imputation ganzer Netze geeignet sind, auf die jeweiligen Inhalte ankommt. Mit der Vorbereitung zahlreicher und inhaltlich unterschiedlicher Datenbereiche ist damit das Problem nicht gelöst. Es sind damit aber Voraussetzungen geschaffen, um solche Probleme mit den hier angebotenen Dateien und Programmen bei geringem Aufwand zielgerichtet anzugehen.

Tabelle 2: Außengrad und Verteilung der Häufigkeit ausgehender Wahlen

Frequency distribution, outgoing links
(Frequencies/ 9999/ Totals of knots and links)
0 1 2 3 4 5 6 7 8
---
z2observed
46 35 41 25 19 10 3 2 1 9999 182 358
25.3 19.2 22.5 13.7 10.4 5.5 1.6 1.1 .5
z2i-nn: all data without network data (1-5)
12 33 76 42 14 3 2 9999 182 394
6.6 18.1 41.8 23.1 7.7 1.6 1.1
z2i-out: only outgoing links
65 55 32 20 10 9999 182 219
35.7 30.2 17.6 11.0 5.5
z2i-po: position census only (from 116 networks in CNETZ)
33 39 50 39 14 5 1 1 9999 182 350
18.1 21.4 27.5 21.4 7.7 2.7 .5 .5

Tabelle 3: Innengrad und Verteilung der Häufigkeit eingehender Wahlen

Frequency distribution, incoming links
(Frequencies/ 9999/ Totals of knots and links)
0 1 2 3 4 5 6 7 8
---
z2observed
36 44 37 36 16 11 1 1 9999 182 358
19.8 24.2 20.3 19.8 8.8 6.0 .5 .5
z2i-nn: all data without network data (1-5)
68 32 18 14 20 9 6 9 3 1 1 0 0
0 0 37.4 17.6 9.9 7.7 11.0 4.9 3.3 4.9 1.6 .5
.5 .0 .0 .0 .0 15 16 17 0 0 1 9999 182
394.0 .0 .5
z2i-out: only outgoing links
83 34 31 22 7 3 1 0 1 9999 182 219
45.6 18.7 17.0 12.1 3.8 1.6 .5 .0 .5
z2i-po: position census only (from 116 networks in CNETZ)
63 30 26 23 16 15 4 4 1 9999 182
350
34.6 16.5 14.3 12.6 8.8 8.2 2.2 2.2 .5

Tabelle 4: Verteilung der Häufigkeit gegenseitiger/symmetrischer Wahlen

Frequency distribution, symmetric links
(Frequencies/ 9999/ Totals of knots and links)
0 1 2 3 4
---
z2observed
82 58 26 14 2 9999 182 160
45.1 31.9 14.3 7.7 1.1
z2i-nn: z2i-nn: all data without network data (1–5)
178 4 9999 182 4
97.8 2.2
z2i-out: only outgoing links
141 29 9 3 9999 182 56
77.5 15.9 4.9 1.6
z2i-po: position census only (from 116 networks in CNETZ)
175 6 1 9999 182 8
96.2 3.3 .5

Tabelle 5: Triplet-Transitivität und Verteilung der Triplet-Häufigkeiten

z2observed
transitive / intransitive triplets: 214 503
Index of transitive triplets: T/(T+I)) .296399
triplet census (1) type no. (2) freq.
1:--- 2:--+ 3:-+- 4:-++ 5:+-- 6:+-+ 7:++- 8:+++

5738407 62990 63163 560 63047 676 503 214
z2i-nn: z2i-nn: all data without network data (1-5)
transitive / intransitive triplets: 44 879
Index of transitive triplets: T/(T+I)) .047619
triplet census (1) type no. (2) freq.
1:--- 2:--+ 3:-+- 4:-++ 5:+-- 6:+-+ 7:++- 8:+++
5720071 68572 68341 1656 69349 648 879 44
z2i-out: only outgoing links
transitive / intransitive triplets: 100 246
Index of transitive triplets: T/(T+I)) .287356
triplet census (1) type no. (2) freq.
1:--- 2:--+ 3:-+- 4:-++ 5:+-- 6:+-+ 7:++- 8:+++
5812274 38792 38750 324 38870 204 246 100
z2i-po: position census only (from 116 networks in CNETZ)
transitive / intransitive triplets: 24 771
Index of transitive triplets: T/(T+I)) .030189
triplet census (1) type no. (2) freq.
1:--- 2:--+ 3:-+- 4:-++ 5:+-- 6:+-+ 7:++- 8:+++
5743031 61324 61203 1002 61555 650 771 24

Zusammenfassung und Ausblick

Dieser Aufsatz beschreibt einen umfangreichen, z.T. auf die Entwicklung sozialer Beziehungsnetze über 9 Wochen bezogenen Datensatz. Es werden die besonderen Probleme diskutiert, die durch fehlende Netzdaten in einzelnen Wochen entstehen. Verfahren der Imputation ganzer Beziehungs-Netze zum Ersatz dieser fehlenden Netzdaten stehen im Mittelpunkt.

Vor einem Ersatz fehlender Daten durch Imputation sollte jedoch die Prüfung ihrer Folgen stehen. Zu entsprechenden Prüfungen werden die verfügbaren – also nicht fehlenden – Daten kurz beschrieben und exemplarisch in 37 Merkmals-Bereichen zusammengefasst. Diese Bereiche können einzeln oder in Kombination zur Suche nach «nächsten Nachbarn» eines Befragten genutzt werden. Soweit die Netzdaten eines Befragten mangels Angaben in einer Woche fehlen, können «nächste Nachbarn» mit ihren verfügbaren Daten zum Ersatz der fehlenden Daten dienen.

Die einzelnen Bereiche verfügbarer Daten werden zur leichten Vergleichbarkeit zu paarweise (formal) gleich normierten Distanzen umgeformt. Damit können diese Bereiche unter Verbindung mit inhaltlich begründeten Gewichten leicht zusammengefasst werden.

Durch konkrete Vergabe solcher Gewichte wurden 3 Bereiche gebildet, auf ihrer Basis eine Imputation aller 116 Netze vorgenommen, und als Beispiel eines dieser Netze mit dem tatsächlich erhobenen Netz verglichen.

Es zeigte sich, dass je nach gewählten Datenbereichen und entsprechend ermittelten «nächsten Nachbarn» die Imputation des gewählten Beziehungsnetze sehr unterschiedlich mit dem entsprechenden beobachteten Netz harmoniert. Bei der Imputation fehlender Beziehungsnetze muss deshalb aufgrund der inhaltlichen Zielsetzung die Auswahl und Gewichtung der Bereiche bedacht werden, die zur Suche nach den «nächsten Nachbarn» dienen sollen.

In diesem Aufsatz werden dazu systematisch nur die verfügbaren Datenbereiche ermittelt und in einheitlich normierter Form als Distanz-Listen dargestellt. Die Imputation und ihre Prüfung bleibt dagegen auf wenige Beispiele beschränkt. Da aber Daten und Programme frei verfügbar sind, bietet sich hier ein breites Feld weiterführender Untersuchungen, die einstweilen vom Autor auch noch beratend unterstützt werden können.

Unter eher statistischen Gesichtspunkten ist die Steuerung der Imputation durch Suche nach bestimmten «nächsten Nachbarn» unbefriedigend, solange nicht Aussagen über die auch zufallskritische Prüfung möglich wird. Dieser Punkt wird im vorliegenden Aufsatz nur am Rande und unter Hinweis auf Beispiele in der Literatur erwähnt. Ebenfalls vorbereitete Programme unterstützen aber auch die zufallskritische Prüfung des hier beschriebenen Verfahrens. An die Stelle der Suche nach «einzelnen (oder nach 3) nächsten Nachbarn» aufgrund geringster Distanz tritt dabei die Suche nach der Verteilung von mehr oder weniger nahen Nachbarn, die entsprechend ihrer relativen Distanz im Simulationsprozess mit unterschiedlicher Häufigkeit ausgewählt werden.

Erzeugt man auf diese Weise eine größere Anzahl (z.B. 1000) imputierter Netze, so kann die Häufigkeitsverteilung ausgewählter Eigenschaften dieser Netze zur Schätzung stochastischer Eigenschaften genutzt werden.

Diese wie die vorher genannten Arbeiten sind hier jedoch nur durch frei verfügbare Distanz-Listen und Programme vorbereitet worden. Sie können in kleinerem wie größerem Umfang zur Durchführung weiterführender Arbeiten genutzt werden, etwa für Arbeiten im Umfang einer Hausarbeit bis zu einer Dissertation.

Список литературы

1. Baltes-Götz, Bernhard 2013: Behandlung fehlender Werte in SPSS und Amos. Mimeo, Zentrum für Informations-, Medien- und Kommunikationstec-hnologie (ZIMK) der Universität Trier.

2. Burt, Ron 1990: Detecting role equivalence. In: Social Networks 12/1990, S. 83–97.

3. Hummell, Hans. J. und Wolfgang Sodeur 1984: Strukturentwicklung unter Studienanfängern. Ein Werkstattbericht. Mimeo, siehe Kapitel 2 unter www.uni-due.de/Sodeur/.

4. Hummell, Hans J. und Wolfgang Sodeur 1987: Triaden- und Tripplettzensus als Mittel der Strukturbeschreibung, S.129-161. In: Franz Urban Pappi 1987.

5. Hummell, Hans J. und Wolfgang Sodeur 1987: Strukturbeschreibung von Positionen in sozialen Beziehungsnetzen, S. 177–202. In: Franz Urban Pappi 1987.

6. Pappi, Franz Urban (Hrsg.) 1987: Methoden der Netzwerkanalyse. Band 1 der Techniken der empirischen Sozialforschung, hrsg. Von Jürgen van Koolwijk und Maria Wieken-Mayser, Oldenbourg, München.

7. Snijders, Tom: Siena bzw. RSiena in der jeweils aktuellen Version, siehe Internet-Seite www.stats.ox.ac.uk/~snijders/siena/.

8. Steindorf, Karen und Oliver Kuß 2011: Multiple Imputation – der State-of-the-Art-Umgang mit fehlenden Werten. Mimeo, Institut für Medizinische Epidemiologie, Biometrie und Informatik an der Martin-Luther-Universität Halle-Wittenberg.

9. Znidarsic, Anja, Patrick Doreian und Anuska Ferligoj 2018: How many nearest neighbours are needed to successfully estimate missing ties due to actor non-response when blockmodelling structure is investigated? (Vortrag auf der Sunbelt Conference 2018 im Juli 2018 in Utrecht).

КОМПЕНСАЦИЯ НЕДОСТАЮЩИХ ДАННЫХ ПРИ ПРОВЕДЕНИИ ПАНЕЛЬНОГО ОПРОСА СТУДЕНТОВ ДВУХ ПЕРВЫХ МЕСЯЦЕВ ОБУЧЕНИЯ

В. Зодер

Университет Дуйсбург-Эссен, Германия

В университете Дуйсбург-Эссен был проведен опрос студентов-первокурсников об их впечатлениях о первых девяти неделях обучения. Респондентам были заданы вопросы об их социально-демографических характеристиках, об отношении к учебным мероприятиям и подразделениям университета, с которыми они успели познакомиться за первые недели обучения, а также о количестве дружеских контактов, которые начинающие студенты успели установить к этому времени. При опросе использовалась случайная выборка при помощи генератора имен.

На отдельных этапах сбора данных была выявлена недостаточность полученной информации, в частности, у ряда студентов (около 200 человек) отсутствовала информация о количестве дружеских контактов, установленных в определенные недели. В таких условиях проведение динамического анализа формирования социальной сети отношений становится невозможным.

В подобных ситуациях применение метода компенсации недостающих данных, как правило, позволяет легко заполнить существующие пробелы. Однако в данном случае мы используем методы компенсации для восполнения определенной части уже полученных данных. Недостающие данные, полученные методом компенсации, затем сравниваются с результатами более поздних замеров. Такой подход позволяет проверить пригодность различных методов компенсации недостающих данных.

Описанный в данной публикации метод компенсации недостающих данных позволяет решать различные задачи. Для этого части полученных данных сначала обрабатываются отдельно, а затем суммируются с использованием соответствующих весовых коэффициентов. Выбор различных весов позволяет адаптировать процедуру компенсации недостающих данных под конкретную цель исследования. Чтобы сделать понимание принципов работы метода компенсации недостающих данных более простым, мы используем только небольшую часть массива данных. Затем мы сравним полученные результаты с результатами более поздних замеров.

Массив данных и AWK-программы находятся в свободном доступе. С ними можно ознакомиться по адресу: Wolfgang.Sodeur@t-online.de.

Результаты сравнительного анализа могут быть загружены со страницы: www.uni-due.de/Sodeur/ вместе с файлом NN18IMP.zip.

Ключевые слова: недостающие данные, панельный опрос, социальная сеть отношений, компенсация, результаты сравнительного анализа.