

Praktische Anwendungsbeispiele zum Umgang mit den pairfam-Daten (Welle 1)

Josef Brüderl, Volker Ludwig, Klaus Pferr,
Nina Schumann

Universität Mannheim
Mai 2011

Gefördert als Langfristvorhaben durch die Deutsche Forschungsgemeinschaft (DFG)

I. Einleitung

Jedes Datenerhebungsprojekt wird letztlich an den Analysen gemessen, die mit den resultierenden Daten tatsächlich realisiert werden. Dies gilt auch für das Beziehungs- und Familienpanel. Der vorliegende Beitrag verfolgt in diesem Zusammenhang vor allem drei Ziele. Erstens soll pairfam-Nutzern der Einstieg in die Datenanalyse erleichtert werden, indem einige Anwendungsbeispiele vorgestellt werden, die in die wesentlichen Eigenschaften der Daten einführen. Zweitens sollen Analysestandards gesetzt werden, indem Lösungen für typische Probleme bei der Datenanalyse aufgezeigt werden. Schließlich soll das Analysepotential der Daten aufgezeigt werden.

Auch wenn erst eine Welle der Daten verfügbar ist, ist die Datenstruktur bereits recht komplex – unter anderem aufgrund des Kohortendesigns, des Multi-Actor-Designs und der retrospektiv erhobenen Biographien. Bislang sind mit der ersten Welle der pairfam-Daten vier Datensätze verfügbar: jeweils ein Datensatz zur Ankerbefragung (anchor1) und zur Partnerbefragung (partner1) sowie zwei generierte Ereignisdatsätze zur Partnerschaftsbiographie der Ankerperson (biopart) und zur Biographie der Kinder der Ankerperson (biochild). Mit der Komplexität der Daten gehen aber auch interessante Anwendungsmöglichkeiten einher. Die Kapitel dieses Sammelbands präsentieren inhaltliche Ergebnisse, die mit diesen Daten (Version 1.0) gewonnen wurden. Der vorliegende Beitrag soll im Gegensatz dazu Anleitung für die Datenanalyse sein – ohne bereits substantielle Forschung betreiben zu wollen.

Zunächst werden wir einige Hinweise allgemeiner Art behandeln, die nach unserer Ansicht die Arbeit mit den Daten erleichtern (Abschnitt II). Anschließend stellen wir anhand konkreter Beispiele vor, welches Potential die pairfam-Daten besitzen, und wie typische Herausforderungen, die im Verlauf eines Aufbereitungs- und Analyseprozesses auftreten, bewältigt werden können. Dabei widmen wir uns folgenden Fragen: Soll eine Gewichtung vorgenommen werden – und wenn ja, welche Verfahren sind dafür angemessen (Abschnitt III)? Wie lassen sich Daten zusammenführen, um etwa dyadische Analysen durchzuführen (Abschnitt IV)? Welche Anwendungs- und Analysemöglichkeiten bieten die retrospektiv erhobenen biographischen Ereignisdaten (Abschnitt V)?

Die Auswertungsbeispiele werden mit dem Statistikprogramm Stata durchgeführt. Es wird davon ausgegangen, dass der Leser mit grundlegenden Syntax-Eigenschaften von Stata vertraut ist (s. Kohler & Kreuter 2008). Aus Platzgründen kann nur eine Auswahl zentraler Aufbereitungs- bzw. Analyseschritte beschrieben werden. Zu den vorgestellten Beispielen sind kommentierte („Quick Start“) Stata do-files auf www.pairfam.de bereitgestellt, die u.a. die hier behandelten Themen in ähnlicher oder identischer Weise behandeln.

II. Allgemeine Hinweise zur Arbeit mit den pairfam-Daten

Auch wenn vieles im Umgang mit den pairfam-Daten letztlich von der konkreten Fragestellung und Analysestrategie abhängen wird, gibt es doch einige Tipps, die die Arbeit mit den Daten grundsätzlich erleichtern können, und die wir unseren Nutzern daher vorab an die Hand geben wollen.

Der erste Hinweis betrifft die Dokumentation der Daten. Die „Quick Start“ do-files, die in den nächsten Abschnitten auszugsweise vorgestellt werden, ermöglichen es dem Nutzer, direkt in die Datenanalyse einzusteigen. Darüber hinaus bieten diese do-files einige nützliche Tipps, u.a. zur Kodierung von Variablen, zur Konstruktion von Stichproben und zu verschiedenen Analysestrategien. Für die tatsächliche Forschung wird selbstverständlich eine genauere Kenntnis des Datengenerierungsprozesses nötig sein. Dazu stehen dem Nutzer verschiedene Quellen zur Verfügung, die mit den Daten ausgeliefert werden und auch online verfügbar sind.

Für detaillierte Informationen zur Datenaufbereitung, unter anderem zur Fehlerbereinigung und zur Erstellung nutzerfreundlicher Variablen und Datensätze, verweisen wir auf das pairfam Data Manual (Brüderl et al. 2010). Der vollständige Wortlaut aller Fragen an die Befragten sowie die resultierenden Variablen mit ihren Ausprägungen sind in den Codebüchern der Anker- und Partnerbefragung dokumentiert. Das Skalenhandbuch (Walper et al. 2010) stellt die im Ankerpersonen-Interview der Welle 1 verwendeten Itemgruppen vor, gibt Auskunft über ihre jeweilige Herkunft sowie ihr theoretisches Konstrukt und enthält Vorschläge zur Bildung von Indizes bzw. Skalen. Informationen zum Feldverlauf der ersten Welle schließlich finden sich im Methodenbericht (Suckow und Schneekloth 2009). Für Fragen, die auch durch die Dokumentation nicht geklärt werden können, steht der pairfam-Nutzerservice zur Verfügung.

Die zweite Empfehlung ist, für eigene Auswertungen wenn möglich auf generierte Variablen bzw. Datensätze zurückzugreifen. Zum einen sind diese generierten Daten so weit aufbereitet, dass sie direkt für Analysen verwendet werden können. Inkonsistenzen zwischen Angaben der Befragten sind hier größtenteils bereinigt. Zudem wurden meist Informationen aus mehreren Variablen kombiniert, so dass dem Nutzer eigene Kodierungen erspart bleiben. Das pairfam-Team beansprucht zwar nicht, die jeweils beste aller möglichen Kodierungen für die generierten Variablen angewendet zu haben. Dennoch erfolgte die Kodierung mit dem Ziel, zunächst einen Standard zu setzen, den die Nutzer übernehmen können, von dem sie aber auch in nachvollziehbarer Weise abweichen können. Zu diesem Zweck wird die Aufbereitung der generierten Daten veröffentlicht. Die entsprechenden Stata do-files (gepackt in der Datei genvarsdo.zip) werden mit den Daten ausgeliefert und sind auch online verfügbar. Der Nutzer kann so die Kodierentscheidungen bei der Erstellung der generierten Daten nachvollziehen. Er kann beschließen, die Kodierung für ihre Zwecke anzupassen. In diesem Fall sollten die selbst vorgenommenen Abweichungen dokumentiert werden. Aus unserer Sicht leisten wir mit der Veröffentlichung dieses Schritts der Datenaufbereitung einen wesentlichen Beitrag zur Verbesserung der Replizierbarkeit und Vergleichbarkeit von Forschungsergebnissen. Nähere Informationen zu generierten Variablen und Datensätzen erhalten Sie im pairfam Data Manual (Brüderl et al. 2010). Zudem stehen generierte Skalenvariablen für Ankerperson und Partner bereit (Walper et

al. 2010). Die jeweilige SPSS-Syntax zur Erzeugung dieser Variablen steht auf der pairfam-Homepage zur Verfügung.

Der dritte Hinweis betrifft eine spezielle Gruppe generierter Variablen („flag-Variablen“), denen besondere Beachtung geschenkt werden sollte, weil sie auf Inkonsistenzen in den Daten verweisen. Bei der Datenaufbereitung wurde auf Rekodierungen von Inkonsistenzen größtenteils verzichtet, da in den meisten Fällen nicht ohne weiteres entschieden werden kann, welche der widersprüchlichen Angaben falsch ist. Deshalb werden mit den Daten Variablen ausgeliefert, die Inkonsistenzen zwischen Variablen anzeigen. Beispielsweise markiert die Variable flag18 Widersprüche zwischen dem selbstberichteten Geschlecht des Partners bzw. der Partnerin und der entsprechenden Proxy-Angabe der Ankerperson. Es liegt in der Verantwortung des Nutzers, solche Inkonsistenzen bei der Auswertung zu berücksichtigen. Wir empfehlen daher, stets die Liste der flag-Variablen auf ihre Relevanz für die jeweilige Analyse hin zu prüfen. Diese Liste finden Sie ebenfalls im pairfam Data Manual. In vielen Fällen wird es sinnvoll sein, inkonsistente Informationen von der Auswertung auszuschließen.

Schließlich wollen wir noch auf wegen der Filterführung fehlende Angaben hinweisen. Bei der Datenaufbereitung wurden in allen Datensätzen einheitliche (durchgängig negative) Werte für missings vergeben. Hat eine Person auf einer Variable den Wert -3 („Trifft nicht zu“), bedeutet das, dass die entsprechende Frage der Person nicht gestellt wurde, weil die Frage für die Person entweder unpassend ist oder weil die Information schon aus vorhergehenden Angaben erschlossen werden kann. Der Nutzer sollte sich deshalb mit der Filterführung vertraut machen, die in den Codebüchern dokumentiert ist. Es empfiehlt sich nicht, Personen mit fehlenden Angaben per *casewise deletion* ohne weitere Prüfung von der Analyse auszuschließen. Man sollte sich darüber im Klaren sein, welche Teilstichprobe man durch dieses Vorgehen ausschließen würde. In vielen Fällen ist es sinnvoll oder sogar notwendig, gültige Werte abzuleiten und fehlende Angaben zu ersetzen.

Grundlegend sei zu guter Letzt der Umgang mit dem Kohortendesign erläutert. pairfam ist als Kohortenstudie angelegt. Damit man mit der Analyse familienrelevanter Prozesse nicht Jahre warten muss, wurde nicht nur eine Kohorte 15-17jähriger gezogen, sondern es wurden zwei ältere Kohorten mitgezogen („Kohortensequenzdesign“ bzw. „Accelerated Longitudinal Design“). Deshalb sollte es der Normalfall sein, die drei Kohorten getrennt zu analysieren (oder nur einzelne Kohorten zu betrachten). Poolt man die Daten der drei Kohorten und führt eine gemeinsame Analyse durch, so ist zu beachten, dass die Grundgesamtheit nicht die deutschsprachige Bevölkerung mit allen Altersgruppen ist. Bei einer gepoolten Analyse mit pairfam-Daten ist die Grundgesamtheit die deutschsprachige Bevölkerung der Jahrgänge 1971-73, 1981-83 und 1991-93.

III. Gewichtungsverfahren

Im Folgenden soll die adäquate Verwendung der im Datensatz verfügbaren Gewichte dargestellt werden. Das hierfür relevante do-file ist „Quick Start.do“.

(1) In gepoolten Analysen muss man gewichten. Die Bruttostichprobengrößen der drei Kohorten wurden so gewählt, dass die Nettostichproben ungefähr gleich groß sind. Da die Größenverhältnisse der drei Kohorten in der Bruttostichprobe nicht den Verhältnissen in der Bevölkerung entsprechen, muss ein Designgewicht verwendet werden, um diese Unterschiede auszugleichen. (2) Weiterhin ist von einem systematischen Nonresponse der Ankerpersonen in der ersten Welle auszugehen, der durch ein Post-Stratification-Gewicht (Anpassungsgewicht) eventuell ausgeglichen werden kann. Da die Verwendung von Post-Stratification-Gewichten stark umstritten ist, überlassen wir dem Nutzer die Entscheidung, das bereitgestellte Gewicht zu verwenden. (3) Wenn der Nutzer sowohl das Design- als auch das Post-Stratification-Gewicht verwenden möchte, sollte er das bereitgestellte Produktgewicht verwenden und nicht selbst das naive Produkt aus beiden Faktoren bilden, da sich dadurch die Fallzahl ändert.

Ein typisches Beispiel für die Verwendung des Post-Stratification-Gewichts ist eine Auszählung der Anzahl der Kinder. Es ist zu erwarten und aus der Literatur bekannt (z.B. Lynn et al. 2005), dass die Anzahl der Kinder einen Einfluss auf die Interviewbereitschaft hat. Daraus resultiert ein selektiver Rücklauf, d.h. die mittlere Anzahl der Kinder in der Stichprobe sollte sich von der mittleren Anzahl in der Grundgesamtheit unterscheiden. Eine Möglichkeit, dies zu kompensieren, bietet das Post-Stratification-Gewicht.

```
replace nkidsbioalv=nkids if nkidsbioalv==-7
mean nkidsbioalv, over(cohort)
mean nkidsbioalv [pweight=psweight], over(cohort)
```

Tabelle 1 zeigt, dass die ungewichtete mittlere Anzahl leiblicher Kinder in der Geburtskohorte 1991-93 ungefähr 0, in Kohorte 1981-83 0,4 und in der Geburtskohorte 1971-73 ungefähr 1,4 ist. Durch die Gewichtung ändern sich diese Zahlen je nach Kohorte mehr oder weniger stark: für die jüngste Kohorte ist die gewichtete mittlere Anzahl immer noch ungefähr 0, für die mittlere Kohorte ungefähr 0,3 und für die älteste Kohorte ungefähr 1,2.

Tabelle 1: Verteilung Kinderzahl nach Kohortenzugehörigkeit

Kohorte	Ungewichtet		Gewichtet (ohne svy-Kommando)		Gewichtet (mit svy-Kommando)	
	Mean	Std. Fehler	Mean	Std. Fehler	Mean	Std. Fehler
1 (1991-1993)	0,001	0,001	0,002	0,001	0,002	0,001
2 (1981-1983)	0,402	0,012	0,320	0,011	0,320	0,011
3 (1971-1973)	1,420	0,018	1,239	0,021	1,239	0,021
N	12.402		12.402		12.402	

Es ist darauf hinzuweisen, dass das Post-Stratification-Gewicht in unserem Beispiel wie ein Designgewicht verwendet wird ([pweight=psweight]). Das ist theoretisch eigentlich nicht adäquat, denn bei dem bereitgestellten Gewicht handelt es sich um ein Raking-Gewicht, d.h. einem Fall kann nicht ein eindeutiges Stratum mit einer Bruttofallzahlgröße in der Grundgesamtheit zugeordnet werden. Dennoch wird dies in der Praxis üblicherweise so gehandhabt.

Im oben dargestellten Beispiel wird die einfache Einbindung von Gewichten direkt in den Befehl verwendet [`pweight=psweight`]. Bei komplizierteren Datenstrukturen empfiehlt Stata die Verwendung der `svy`-Kommandos, die eine volle Spezifikation des Auswahlprozesses erlauben. Eine ausführliche Darstellung der Kommandos findet sich bei StataCorp. (2009) und Heeringa et al. (2010). Im Beispiel sind die Ergebnisse beider Vorgehensweisen identisch (vgl. Tabelle 1).

```
svyset [pweight=psweight]
svy: mean nkidsbioalv, over(cohort)
```

Für die Schätzung der gewichteten Anteile einer kategorialen Variablen in der Population ist der `mean`-Befehl nicht hilfreich. Je nach gewünschter Detailfülle empfiehlt sich die Verwendung der Befehle `proportion` oder `tabulate`. Der Befehl `proportion` liefert neben den Anteilen in der Population auch die zugehörigen Konfidenzintervalle. Beispielsweise würde man die ungewichtete Verteilung von Familienstand und Beziehungsstatus (Variable „`relstat`“) in der ältesten Kohorte wie folgt berechnen (vgl. Tabelle 2).

```
recode relstat -7=.
proportion relstat if cohort==3
```

Allerdings können die Standardfehler, die man mit `proportion` ausgewiesen bekommt, falsch sein. Um die korrekten Standardfehler zu erhalten, verwendet man die Option `subpop` des `svy`-Kommandos (Tabelle 2).

```
svy, subpop(if cohort==3): proportion relstat
```

Tabelle 2: Verteilung Familienstand und Beziehungsstatus

Familienstand & Beziehungsstatus	Ungewichtet		Gewichtet	
	Anteil	Std. Fehler	Anteil	Std. Fehler
Nie verheiratet gewesen				
Single	0,113	0,005	0,153	0,007
Partnerschaft, LAT	0,043	0,003	0,053	0,004
Partnerschaft, Kohab.	0,098	0,005	0,128	0,007
Verheiratet				
Kohabitation	0,626	0,008	0,573	0,009
LAT	0,006	0,001	0,006	0,001
Geschieden / Getrennt				
Single	0,057	0,004	0,041	0,003
Partnerschaft, LAT	0,022	0,002	0,017	0,002
Partnerschaft, Kohab.	0,032	0,003	0,027	0,003
Verwitwet				
Single	0,002	0,001	0,001	0,001
Partnerschaft, LAT	0,001	0,000	0,001	0,001
Partnerschaft, Kohab.	0,001	0,000	0,001	0,000
N	4.046		4.043	

Auch bei multivariaten Regressionsmodellen können Anpassungsgewichte verwendet werden. Die Verwendung der Gewichte verläuft in Stata analog wie oben dargestellt. Zur Illustration betrachte man die im Folgenden dargestellte OLS-Regression der Kinderzahl auf das Geschlecht und die Bildungsjahre (Tabelle 3).

```
gen woman = sex==2
recode yeduc -7/0=.
reg nkidsbioalv woman yeduc if cohort==3
svy, subpop(if cohort==3): reg nkidsbioalv woman yeduc
```

Tabelle 3: Kinderzahl in Abhängigkeit von Geschlecht und Bildungsjahren (OLS-Regression)

Abh. Variable:	Ungewichtet		Gewichtet	
	Koeffizient	t-Wert	Koeffizient	t-Wert
Kinderzahl				
Geschlecht (Frau)	0,411***	11,69	0,345***	8,35
Bildungsjahre	-0,047***	-7,94	-0,053***	-7,61
Konstante	1,809***	22,13	1,777***	18,34
N	4.033		4.034	
R ²	0,048		0,043	

*** p<0.001

In den bisherigen Beispielen haben wir auf eine gepoolte Analyse über die Kohorten verzichtet, so dass eine Verwendung von Designgewichten nicht notwendig war. Abschließend soll nun noch ein Regressionsbeispiel dargestellt werden, bei dem über die drei Geburtskohorten gepoolt wird. Ein klassisches Beispiel hierfür ist die Bildungsrendite, d.h. eine OLS-Regression des persönlichen Nettoeinkommens auf die Bildungsjahre, hier unter Kontrolle des Alters. Da die Analysen nicht getrennt nach Geburtskohorten durchgeführt werden, muss das Designgewicht verwendet werden (Tabelle 4).

```
svyset [pweight=dxpsweight]
recode incnet -7/0=.
reg incnet yeduc age
svy: reg incnet yeduc age
```

Tabelle 4: Nettoeinkommen in Abhängigkeit von Bildungsjahren und Alter (OLS-Regression)

Abh. Variable:	Ungewichtet		Gewichtet	
	Koeffizient	t-Wert	Koeffizient	t-Wert
Nettoeinkommen				
Bildungsjahre	93,555***	23,72	95,943***	20,27
Alter	34,592***	19,17	41,392***	27,31
Konstante	-929,662***	-14,19	-1140,137***	-17,98
N	5.636		5.214	
R ²	0,181		0,241	

*** p<0.001

Es ist zu beachten, dass in diesem Beispiel beide Gewichte, d.h. das Designgewicht und das Post-Stratification-Gewicht verwendet werden. Bei diesem Vorgehen muss der bereitgestellte kombinierte Faktor `dxpsweight` verwendet werden. Wenn man auf das Post-Stratification-Gewicht verzichten möchte, muss man auf das reine Designgewicht `dweight` zurückgreifen. Bei gepoolten Analysen mit den `pairfam`-Daten muss man sich aber immer der eingeschränkten Aussagekraft bewusst bleiben, denn die Grundgesamtheit ist nicht die deutschsprachige Bevölkerung über alle Altersstufen, sondern nur die deutschsprachige Bevölkerung der drei `pairfam`-Kohorten (s.o.).

IV. Zusammenspielen von Datensätzen und Analysen mit Dyaden

Im folgenden Abschnitt wird demonstriert, wie verschiedene Datensätze – z.B. der Ankerperson und ihres Partners – für Analysen zusammengeführt werden können. Zudem werden eine Datenaufbereitung und ein Auswertungsbeispiel mit Daten über Dyaden dargestellt. Das relevante `do-file` ist „Quick Start.do“. Der Einfachheit halber werden hier und im folgenden Abschnitt keine Gewichtungsstrategien angewandt.

1. Zusammenführen der Daten der Ankerperson und ihres Partners / ihrer Partnerin

In der ersten `pairfam`-Welle wurden Interviews mit den Ankerpersonen sowie – falls zutreffend – mit deren aktuellen Partnern geführt. Es liegen sowohl Proxy-Angaben der Ankerperson über den aktuellen Partner sowie Angaben des Partners selbst (zu teilweise identischen Fragen) vor. Die entsprechenden Informationen wurden in den Datensätzen `anchor1` und `partner1` abgelegt.

Im Statistikprogramm Stata werden verschiedene Datensätze mithilfe des Befehls `merge` zusammengeführt. Ist der Ankerdatensatz geladen, kann – wie im folgenden Beispiel demonstriert – der Partnerdatensatz zugespield werden, indem dieser als „using-Datensatz“ festgelegt wird:

```
merge 1:1 id using partner1.dta
```

Zur Identifizierung des Partners einer bestimmten Ankerperson, d.h. um die richtigen Zeilen der beiden Datensätze einander zuzuspielen, wurde hier die Personennummer der Ankerperson (Variable „`id`“) benutzt, die in beiden Datensätzen enthalten ist. Anschließend werden nur solche Beobachtungen beibehalten, bei denen auch Informationen aus der Partnerbefragung vorliegen (Ausprägung 3 auf der generierten Variable „`_merge`“):

```
keep if _merge==3
```

Während der Erhebung wurde so weit als möglich sichergestellt, dass der schriftlich befragte Partner dem von der Ankerperson genannten aktuellen Partner entspricht. Dennoch sind Abweichungen bzw. Identifizierungsfehler möglich. Es wird daher empfohlen, bei jeder gemeinsamen Analyse von Partner- und Ankerdaten alle Partnerinformationen zu löschen, bei denen – mehr oder weniger starke – Abweichungen bei Geschlecht und Geburtsjahr zwischen den Proxy-

Angaben der Ankerperson und den Antworten seines/ihres Partners bestehen. Dabei steht für Abweichungen des Geschlechts des Partners die generierte Variable „flag18“ zur Verfügung, deren Ausprägungen 1 und 2 eine Inkonsistenz zwischen Partner- und Ankerangaben anzeigen. Folglich werden nur solche Fälle in dem zusammengeführten Datensatz beibehalten, die auf dieser Variablen den Wert 0 aufweisen. Um unplausible Abweichungen beim Geburtsjahr zu identifizieren, wird eine Hilfsvariable generiert, die die Differenz zwischen den beiden Angaben (sofern gültige Werte vorliegen) misst. Nur die Fälle, bei denen die (absolute) Differenz kleiner als 2 Jahre ist, verbleiben im Datensatz.

```
keep if flag18==0
gen diff = pdoby-sdply if pdoby>0 & sdply>0
keep if abs(diff)<2 | diff==.
```

Generell liefern die zur Verfügung stehenden „flag“-Variablen Hinweise auf weitere inkonsistente Fälle. Im vorliegenden Beispiel zeigt die Variable „flag12“ weitere Inkonsistenzen hinsichtlich des Geburtsjahrs des Partners an. Auch hier ist es ratsam, nur Fälle beizubehalten, die keine der definierten unplausiblen Fälle darstellen.

```
keep if flag12==0
```

Nun können Analysen mit dem zusammengeführten und bereinigten Datensatz durchgeführt werden.

2. Analysen mit Dyaden: Ankerperson und Eltern

Die Eltern der Ankerperson wurden in der ersten pairfam-Befragungswelle noch nicht selbst befragt (dies geschieht zum ersten Mal in der zweiten Welle). Es sind jedoch Angaben der Ankerpersonen zu ihren Eltern vorhanden. Analysen von Eltern-Kind-Dyaden sind so bereits mit den Daten der ersten Welle möglich. Im Folgenden soll anhand eines einfachen Beispiels veranschaulicht werden, welchen Einfluss das Alter und das Geschlecht des Ankers bzw. des Elternteils auf die Verbundenheit der Dyade (aus Sicht des Ankers) haben¹.

Zunächst wird aus den vorliegenden Informationen der Ankerbefragung die abhängige Variable generiert: Verbundenheit mit dem Vater („close1“) sowie mit der Mutter („close2“). Die Variablen igr11 und igr13 bilden hierfür die Ausgangsbasis; die Frage lautete: „Wie eng fühlen Sie sich mit Ihrer leiblichen Mutter [Ihrem leiblichen Vater] heute verbunden?“. Die Antwortskala reichte von 1 „überhaupt nicht eng“ bis 5 „sehr eng“. Die Frage wurde nur dann gestellt, wenn Kontakt zu dem entsprechenden Elternteil bestand, d.h. wenn Vater bzw. Mutter nicht verstorben sind und die Ankerperson Kontakt zu diesem Elternteil hat. Den neu generierten Variablen close1 und close2 wird jeweils der Wert 0 zugeordnet, wenn Vater bzw. Mutter verstorben sind.

¹ Im Stata do-file „Quick-Start.do“ ist ein ähnliches Beispiel mit der zu erklärenden Variable „Kontakt zu den Eltern“ enthalten.

```

gen      close2=igr11 if inrange(igr11,1,5)
replace close2=0 if igr5==2
gen      close1=igr13 if inrange(igr13,1,5)
replace close1=0 if igr7==2

```

Analog werden die Variablen „agepar1“ (Alter Vater) und „agepar2“ (Alter Mutter) aus den Variablen „fage“ und „mage“ gebildet. Das Geschlecht von Mutter und Vater ergibt sich automatisch (s.u.). Alter und Geschlecht der Ankerperson sind Standardvariablen.

Mithilfe des `reshape`-Befehls wird nun aus dem bisherigen „wide“-Format ein „long“-Format gebildet. Im „wide“-Format stehen die Informationen über die Beziehung zu Vater und Mutter auf einer Datenzeile. Im „long“-Format soll jede Anker-Elternteil-Dyade eine eigene Datenzeile haben. Generell überführt der Befehl `reshape long` einen Datensatz aus dem wide-Format in einen Datensatz mit long-Format:

```

reshape long close@ agepar@, i(id) j(parent)

```

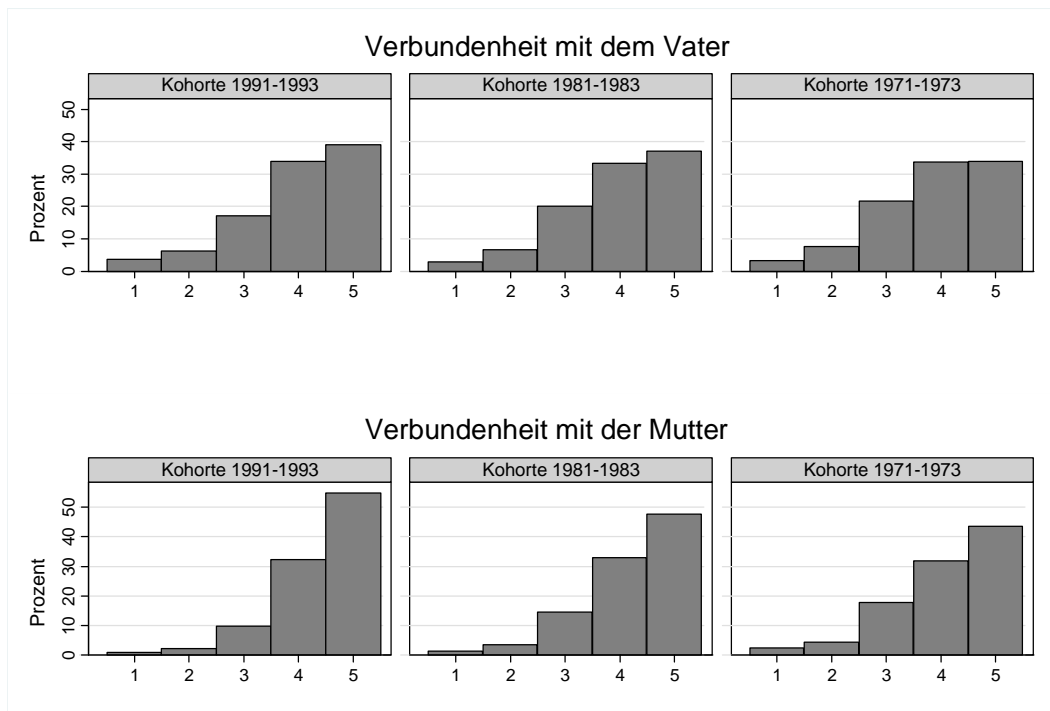
Die Variable, die in der Klammer hinter „i“ angegeben ist, ist die Personnummer der Ankerperson. Die in der Klammer nach „j“ angegebene Variable wird neu generiert und identifiziert die Art der Dyade (1: Anker-Vater, 2: Anker-Mutter). Die Werte der j-Variable werden aus den Dyadenvariablen entnommen, wobei „@“ als Platzhalter angibt, wo der Wert steht. Die neu entstandene Variable „parent“ zeigt dann an, ob die jeweilige Dyade aus Ankerperson und Mutter („2“) oder Ankerperson und Vater („1“) besteht.

Der so erstellte Datensatz im long-Format enthält nun für jeden Anker zwei Datenzeilen: jeweils eine für die Anker-Vater- und die Anker-Mutter-Dyade. In jeder Datenzeile sind Alter und Geschlecht des Ankers und des jeweiligen Elternteils vermerkt sowie die Verbundenheit des Ankers mit dem jeweiligen Elternteil. Es müssen nun noch die nicht relevanten Dyaden gelöscht werden. Hier sind das die Dyaden, bei denen das Elternteil bereits verstorben ist oder kein Kontakt mehr besteht².

Mit dem generierten Dyaden-Datensatz kann man nun Analysen über Dyaden machen. Um einen Einblick in die Verteilung der abhängigen Variable „Verbundenheit mit den Eltern“ zu erlangen, zeigen die Histogramme in Abbildung 1 die Verteilung der Antworten differenziert nach Elternteil und Kohorten. Es wird deutlich, dass die sehr starke Verbundenheit mit den Eltern (Ausprägung 5) über die Geburtskohorten abnimmt, während v.a. bei der Mittelkategorie ein gegenläufiges Muster deutlich wird.

² Da für die Dyaden ohne Kontakt (N=1.184) aus erhebungstechnischen Gründen nicht nach der Verbundenheit gefragt wurde, müssen sie aus der Analyse ausgeschlossen werden.

Abbildung 1: Verbundenheit mit Vater bzw. Mutter nach Kohortenzugehörigkeit



Die Verbundenheit mit den Eltern soll nun multivariat untersucht werden. Aufgrund des Skalenniveaus der abhängigen Variablen werden ordinale logistische Regressionen berechnet.

```
ologit close ageanchor agepar anchorfemale parentmother, vce(cluster id)
```

Die Variable "close" misst die Verbundenheit, "ageanchor" und "agepar" sind die Altersvariablen und „anchorfemale“ und „parentmother“ messen das Geschlecht. Mit der vce-Option wird bei der Berechnung der Standardfehler die Clustering in den Dyaden-Daten berücksichtigt.

Wie aus Tabelle 5 ersichtlich, nimmt mit zunehmendem Alter der Ankerperson die Verbundenheit mit den Eltern ab. Da dies eine Querschnittsanalyse ist, ist allerdings unklar, ob dies ein Kohorten- oder ein Alterseffekt ist. Erst die Analysen mit weiteren Wellen werden dies klären können. Das Alter des Elternteils dagegen scheint keinen Effekt zu haben. Frauen fühlen sich generell ihren Eltern mehr verbunden. Die Verbundenheit mit der Mutter ist besonders hoch.

Tabelle 5: Verbundenheit mit den Eltern in Abhängigkeit von Alter und Geschlecht (ordinale logistische Regression)

	Koeffizient	z-Wert
Alter Ankerperson	-0,024***	-7,34
Alter Elternteil	0,001	0,21
Geschlecht Ankerperson (Frau)	0,204***	6,17
Geschlecht Elternteil (Mutter)	0,550***	25,35
N	18.877	
Pseudo R ²	0,014	

*** p<0.001

V. Analysen mit den Biographiedaten: Sequenz- und Ereignisdatenanalysen

Der Fragebogen der ersten pairfam-Welle beinhaltet detaillierte Fragen zur Partnerschafts-, Ko-habitations- und Heiratsgeschichte der befragten Männer und Frauen ab dem Alter von 14 Jahren sowie zu eigenen Kindern und Episoden des Zusammenlebens mit diesen Kindern. Diese retrospektiv erhobenen biographischen Informationen wurden nutzerfreundlich aufbereitet (u.a. unter Anwendung von Konsistenzchecks) und als zwei zusätzliche Episodendatensätze zugänglich gemacht: „biopart“ sowie „biochild“.

Kapitel 8 des „Data Manual“ (Brüderl et al. 2010) gibt einen eingehenden Überblick über die Episodendatensätze und deren Format, die darin enthaltenen Variablen und ihre Eigenschaften (z.B. Datumsformat) sowie den Umgang mit fehlenden Werten, Zensierungen und Inkonsistenzen. An dieser Stelle soll daher auf eine detaillierte Beschreibung der Datensätze verzichtet werden.

Im Folgenden wird am Beispiel der Partnerschaftsbiographie, die im Datensatz „biopart“ abgebildet ist, die Aufbereitung der Daten für eine Sequenz- und Ereignisdatenanalyse vorgestellt. Die Kommandos findet man in „Quick Start biopart 1.do“ bzw. „Quick Start biopart 2.do“³.

1. Sequenzdatenanalyse

Zunächst soll die Vorgehensweise bei einer Sequenzdatenanalyse vorgestellt werden. Partnerschaftsverläufe – und Lebensverläufe im Allgemeinen – lassen sich als Sequenzen konzipieren, die aus einer definierten Anzahl von in einer spezifischen Weise geordneten Elementen bestehen (Scherer und Brüderl 2010; Brzinsky-Fay et al. 2006). Drei Zielsetzungen sind in diesem Kontext zu differenzieren: Beschreibung der Sequenzen, Vergleich der Sequenzen hinsichtlich ihrer (Un-)Ähnlichkeit und Typisierung sowie Verwendung der Sequenztypen als erklärende Variablen in statistischen Modellen (vgl. Brzinsky-Fay et al. 2006). Dabei kommen sowohl deskriptive Verfahren als auch „Optimal Matching“-Methoden zum Einsatz.

Im vorliegenden Datensatz kann jedem Befragten eine Sequenz zugeordnet werden, die die individuelle Partnerschaftsgeschichte über die Zeit abbildet. Für jeden Monat ab dem 10. Geburtstag wird festgehalten, in wie vielen Partnerschaften sich eine Person befand⁴. Der „Zustand“ 0 bedeutet mithin, dass die Person Single war. Werte größer 1 sind möglich, da multiple Partnerschaften zugelassen werden und auch angegeben wurden.

Zunächst muss die Definition der Zeitachse durch eine Umwandlung der Datumsvariablen in Prozesszeit (mit Eins als Monat des 10. Geburtstages) erfolgen. In den Episodendatensätzen „biopart“ und „biochild“ sind alle Datumsvariablen in Monaten seit Januar 1900 kodiert. Deshalb muss man von allen Datumsangaben (Interviewdatum, Beginn und Ende von Partnerschaften) das Geburtsdatum+119 (Monate vor dem 10. Geburtstag) abziehen.

³ Auch zur Verwendung des Datensatzes „biochild“ steht ein Quick Start do-file auf der Homepage bereit, der hier jedoch nicht näher erläutert wird („Quick Start biochild.do“).

⁴ Auf eine weitere Differenzierung von Partnerschaftsformen anhand des Kohabitationsstatus oder des Familienstands wird hier der Übersichtlichkeit halber verzichtet.

```

gen tint = intdat - (dob + 119)

forvalues x=0/12 {
    quietly replace begp`x`=begp`x'-(dob+119) if begp`x'!=-999
    quietly replace endp`x`=endp`x'-(dob+119) if endp`x'!=-999
}

```

Die Variable „tint“ misst das Datum des Interviews in Monaten seit dem 10. Geburtstag. Das Maximum in den Daten ist 346 Monate, was somit die maximale Länge der Sequenzen ist. Auf eine Standardisierung der Beobachtungsdauer wird verzichtet, weshalb Sequenzen von unterschiedlicher Länge vorliegen. Befragte, die bereits vor dem 10. Geburtstag eine Partnerschaftsepisode angegeben haben, werden aus Plausibilitätsgründen aus dem Datensatz gelöscht.

Nachdem sichergestellt wurde, dass alle Partnerschaftsepisoden positive Dauern aufweisen, folgt das Generieren der zentralen Variablen „inpx“ (mit x von 1 bis höchstens 346), die die Anzahl der von der Ankerperson berichteten Partnerschaften im jeweiligen Monat x angibt. Zunächst wird vom Zustand „Single“ (0) ausgegangen, wobei sich der Wert der Variable anschließend um eins erhöht, wenn der jeweilige Monat x zwischen Beginn und Ende einer Beziehung liegt. Nach dem letzten beobachteten Monat wird die Variable als missing definiert.

```

forvalues x=1/346 {
    gen inp`x`=0
}

forvalues x=1/346 {
    forvalues y=0/12 {
        quietly replace inp`x' = inp`x'+1 if `x'>=begp`y' &
            `x'<=endp`y'
    }
    quietly replace inp`x'=. if `x'>tint
}

```

Ein Aspekt, der die pairfam-Daten zur Partnerschaftsgeschichte von Daten anderer Projekte unterscheidet, ist die explizite Erhebung von Unterbrechungen innerhalb der Partnerschaften. Über 900 Beziehungen weisen mindestens eine Unterbrechung auf. Dieser Tatsache wird Rechnung getragen, indem bei der Bildung der Partnerschaftssequenzen Unterbrechungen berücksichtigt werden. Während einer Unterbrechung wird von „inpx“ eins abgezogen (s. „Quick Start biopart 1.do“ für die Details).

Nachdem die Sequenzen der Partnerschaftsverläufe konstruiert wurden, kann nun mit einer Sequenzdatenanalyse begonnen werden. Eine detaillierte Darstellung der Verfahren und der zugrundeliegenden theoretischen Überlegungen ist im Rahmen dieses Beitrags nicht möglich; der Leser sei auf weiterführende Literatur (Scherer 2001; Scherer und Brüderl 2010; Brzinsky-Fay et al. 2006; Brüderl und Scherer 2004) verwiesen.

Exemplarisch sollen Sequenz-Indexplots erzeugt werden. Sie klassifizieren verschiedene Typen von Beziehungsverläufen, und stapeln sie in einer Abbildung übereinander. Dabei entspricht jede horizontale Linie einer Sequenz, wobei die Zustände durch unterschiedliche Farbabstufungen differenziert werden. Dunkelgrau ist der Zustand „Single“ markiert, während hellgraue Abschnitte den Zustand „in Beziehung“ darstellen. Nach Anzahl der jeweils angegebenen, parallelen Partnerschaften wird nicht mehr differenziert. Die horizontale x-Achse entspricht der Prozesszeit; die vertikale y-Achse gibt die Anzahl der Sequenzen an.

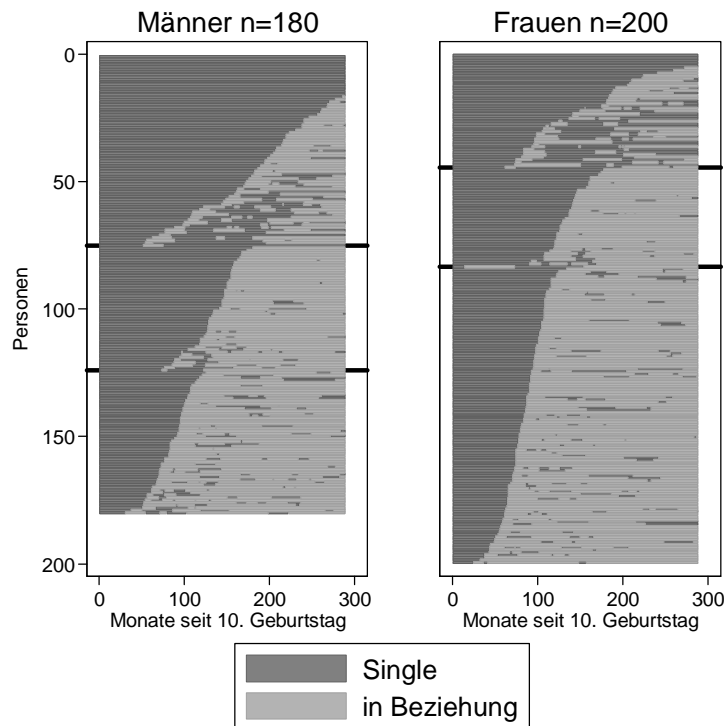
Um die notwendigen Befehle ausführen zu können, muss zunächst das Stata ado-Paket „SQ“ des Autorenteams um Ulrich Kohler installiert werden (vgl. Brzinsky-Fay et al. 2006):

```
ssc install sq
```

Aus dem bisher generierten Datensatz wird eine zufällige 10 % Stichprobe gezogen (Befehl `sample 10`), um eine sinnvolle Darstellung zu ermöglichen (die Anzahl der Sequenzen ist zu groß). Es werden nur die Sequenzen der ältesten Kohorte (35 bis 37 Jahre alt) berücksichtigt. Für die folgenden Analysen ist es sinnvoll die Sequenzlängen zu standardisieren, weshalb die Sequenzen nach dem Alter 33 (Monat 288) abgeschnitten werden.

Um die Sequenzen sinnvoll zu sortieren, wird das Optimal-Matching-Verfahren eingesetzt. Der Befehl `sqom` verwendet den Needleman-Wunsch Algorithmus, um für jedes Sequenzpaar die Levensthein-Distanz und damit ihre Unähnlichkeit zu identifizieren (Brzinsky-Fay et al. 2006; Brüderl und Scherer 2004; Scherer und Brüderl 2010). Die dadurch gewonnene Distanzmatrix wird für eine Clusteranalyse zur Klassifizierung genutzt (Befehl `sqclusterdat`). Dabei wird das Ward-Verfahren gewählt, weil es auf Basis der Levensthein-Distanzmatrix tendenziell homogene Cluster liefert. Die Anzahl der Cluster wird mithilfe eines Dendrogramms festgelegt. Ausgewählt wird die Version mit drei Clustern. Anschließend werden die Clustervariablen zu den Sequenzdaten hinzugefügt und es wird ein Sequenz-Indexplot – getrennt nach Geschlecht – generiert (Abbildung 2).

Abbildung 2: Sequenzindexplots mit Optimal-Matching getrennt für Männer und Frauen;
10%-Zufallsstichprobe



In Abbildung 2 sind die Sequenzen entsprechend ihrer Clusterzugehörigkeit sortiert, wobei die drei verschiedenen Cluster durch horizontale Linien gekennzeichnet sind. Inhaltlich ist das oberste Cluster überwiegend durch den Status „Single“ gekennzeichnet, während die Personen des unteren Clusters überwiegend in einer Beziehung gelebt haben bzw. leben. Auffällig dabei sind die deutlichen Geschlechterunterschiede in der anteilmäßigen Bedeutung der Cluster. Frauen sind deutlich seltener als Männer im oberen Cluster zu finden; mehr als die Hälfte der Frauen sind im Cluster „überwiegend in Beziehung“ zu verorten. Bei Männern ist hingegen das oberste Cluster am häufigsten.

2. Ereignisdatenanalyse

Gegenstandsbereich der Ereignisdatenanalyse ist die Analyse der Dauer, bis ein Ereignis bzw. ein Übergang eintritt. Dabei ist sowohl die Beschreibung der Dauer von Interesse als auch die Betrachtung von Faktoren, die diesen Übergangsprozess beeinflussen (Blossfeld et al. 2007); (vgl. auch Cleves et al. 2008).

Im Folgenden wird eine Ereignisdatenanalyse am Beispiel der Dauer von Beziehungen präsentiert (vgl. do-file „Quick Start biopart 2.do“). Dabei werden zunächst Beziehungsunterbrechungen ignoriert; in einem zweiten Schritt werden diese Unterbrechungen bei der Datenaufbereitung explizit berücksichtigt. Zunächst wird eine Variable generiert, die angibt, wie viele Partnerschaften eine Person insgesamt angegeben hat (maximal zwölf Partnerschaften kamen vor).

```

gen nunions=0

forvalues x=0/12 {
    quietly replace nunions=nunions+1 if indexp`x'>0
}

```

Die Variable „indexp“ wurde bei der Erstellung des Episodendatensatzes „biopart“ generiert und gibt die chronologisch korrekte Reihenfolge der Partnerschaften an (gültige Werte von 1 bis 13, vgl. Brüderl et al. 2010). Anschließend wird aus dem bisherigen „wide-format“ (eine Zeile pro Befragungsperson) ein „long-format“ mit einer Zeile pro Partnerschaft generiert. Dies geschieht mithilfe des `reshape`-Kommandos. Nur Befragte, die mindestens eine Beziehung erlebt haben, werden hierbei und im Folgenden als Risikopopulation noch berücksichtigt. Nach diesen vorbereitenden Schritten werden nun die Daten so organisiert, dass eine Ereignisdatenanalyse möglich wird. Dies umfasst, dass zensierte Beziehungen als Ende das Interviewdatum zugewiesen bekommen und Partnerschaften mit fehlenden Datumsangaben zu Beginn oder Ende gelöscht werden.

Anschließend wird eine Variable gebildet, die die Dauer der jeweiligen Partnerschaft angibt. Bei Dauern von 0, die durch gleiches Beginn- und Enddatum zustande kommen, wird angenommen, dass die jeweilige Beziehung einen halben Monat gedauert hat.

```

gen dur=endp-begp
recode dur 0 = 0.5

```

Um das Ende einer Beziehung zu markieren, wird die Ereignisvariable „separ“ gebildet, die den Wert 1 annimmt, wenn eine Trennung bzw. Scheidung stattfand.

```

gen separ=0

replace separ=1 if reasendp==1 | reasendp==2

```

Mithilfe des Befehls `stset` werden die Daten als Ereignisdaten definiert, wobei die Variable „dur“ die Prozesszeit (hier: die Partnerschaftsdauer) angibt und „separ“ das Ereignis (die Trennung).

```

stset dur, failure(separ)

```

Anschließend stellen wir die Trennungsrate für die Prozesszeit 0 bis 120 (die ersten zehn Jahre einer Beziehung) dar.

```

sts graph, xlabel(0 12 to 120) tmax(120) hazard width(7)

```

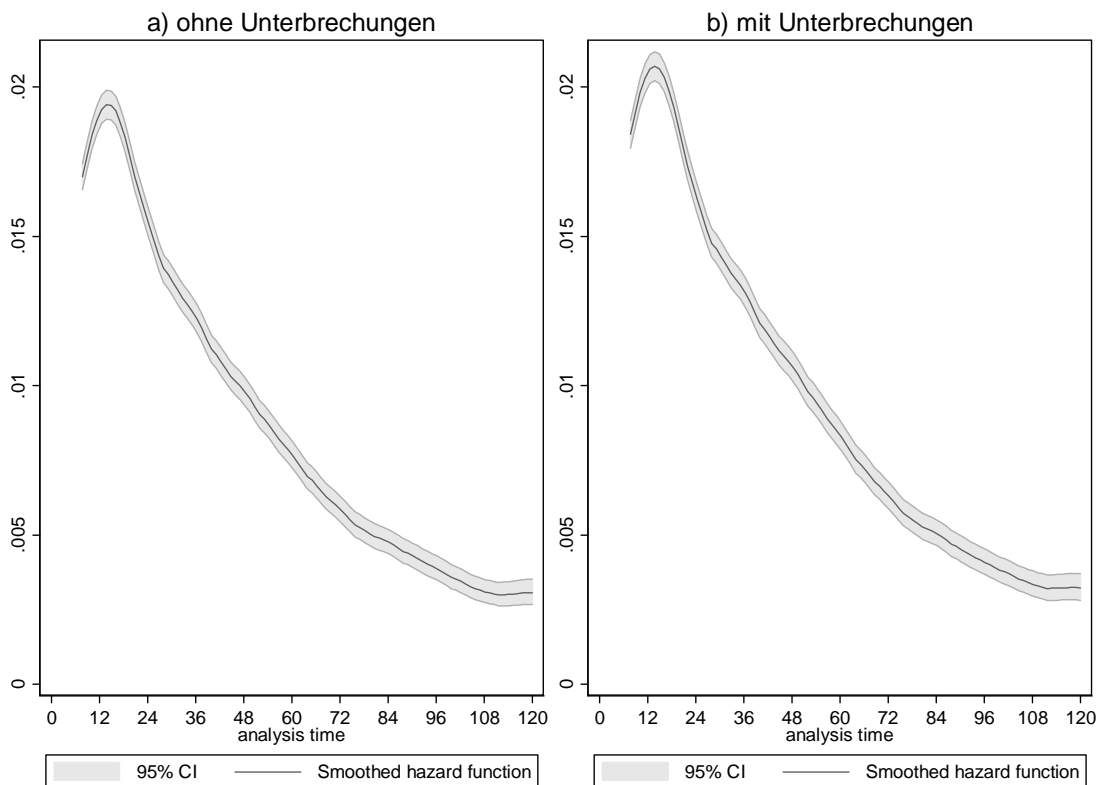
Die linke Seite der Abbildung 3 entspricht dem durch diesen Befehl produzierten Graphen.

Bis hierhin wurde lediglich der (erste) Beginn und das (letzte) Ende einer Partnerschaft berücksichtigt. Wie bereits bei obiger Sequenzdatenanalyse deutlich wurde, besteht eine Besonderheit der `pairfam`-Daten hinsichtlich der Beziehungsgeschichte der Befragten allerdings darin, dass auch Unterbrechungen innerhalb einer Partnerschaft erhoben werden. Als Vorgehensweise wird

hier vorgeschlagen, den Beginn einer Unterbrechung wie eine Trennung und das Ende einer Unterbrechung wie den Beginn einer neuen Beziehung zu behandeln. Somit werden mehrfache Episoden für eine Partnerschaft mit Unterbrechung(en) generiert. Die so erzeugte Trennungsrate ist im rechten Teil der Abbildung 3 festgehalten. Das genaue Vorgehen ist in „Quick Start biopart 2.do“ enthalten.

Abbildung 3 zeigt, dass die Trennungsrate mit zunehmender Prozesszeit abnimmt (nachdem sie im ersten Jahr allerdings ansteigt). Eine Trennung wird mit steigender Beziehungsdauer immer unwahrscheinlicher. Vergleicht man nun die Ergebnisse ohne und mit Berücksichtigung der Unterbrechungen, so wird deutlich, dass sich erwartungsgemäß die Beziehungsdauern verkürzen, wenn Unterbrechungen einbezogen werden. Die Trennungsrate ist insbesondere im ersten Jahr höher, wenn man Unterbrechungen wie Trennungen behandelt.

Abbildung 3: Trennungsraten ohne und mit Berücksichtigung von Unterbrechungen



VI. Diskussion & Ausblick

Ziel dieses Beitrags war, den pairfam-Nutzern die Daten der ersten Erhebungswelle näherzubringen. Es wurden konkrete Anwendungsbeispiele ausgewählt, um Lösungen für gängige Probleme der Datenaufbereitung und -analyse aufzuzeigen. Diskutiert wurde insbesondere die Verwendung von Gewichten, das Matchen von Informationen aus unterschiedlichen Datensätzen und der Umgang mit biographischen Ereignisdaten zur Partnerschaftsgeschichte.

Mit der Datenauslieferung der zweiten Welle wird es in der Arbeit mit den Daten drei wesentliche Änderungen geben. Zum einen werden ab Welle 2 natürlich Paneldaten vorliegen. Zweitens

werden zusätzlich zu den Daten der Ankerpersonen und ihrer Partner auch Daten zu den Eltern und Kindern der Ankerpersonen vorliegen. Und drittens werden weitere nutzerfreundlich aufbereitete Datensätze zur Verfügung gestellt. Mit diesen Änderungen steigt die Komplexität der pairfam-Daten, aber auch das Analysepotential.

Die Tatsache, dass in Welle 2 erstmals Paneldaten vorliegen, wird vor allem eine solidere Analyse kausaler Mechanismen erlauben. Dies setzt jedoch nicht nur Kenntnisse der entsprechenden statistischen Verfahren voraus, sondern auch Kenntnisse im Umgang mit der Datenstruktur. Eine zusätzliche Herausforderung neben dem Matchen von Informationen aus unterschiedlichen Datensätzen ist dann etwa das Matchen von Informationen über die Wellen. Zudem treten dann Fragen zum Umgang mit dem Problem der Panelmortalität und damit zur Längsschnittgewichtung auf.

Mit der Verfügbarkeit der Eltern- und Kinderdaten werden ab Welle 2 komplexere Datenstrukturen denkbar. So werden nicht nur Analysen der Anker-Partner-Dyade, sondern sogar triadische Analysen möglich sein (etwa der Triade Anker-Partner-Kind). Die Analyse der Anker-Partner-Beziehung wird dann erstmals auch im Längsschnitt möglich sein. Mit Blick auf die mögliche Selektivität der Daten aus den Multi-Actor-Studien aufgrund der Rekrutierung der Befragungspersonen durch Zustimmung der Ankerperson (s. Kapitel 1) stellt sich hier ebenfalls die Frage nach Gewichtungstrategien.

Mit den Daten der zweiten Welle werden im Rahmen von pairfam erstmals prospektiv erhobene Paneldaten zur Partnerschafts- und Familienentwicklung in Deutschland vorliegen. Der Erfolg des Datenerhebungsprojekts pairfam wird letztlich an den Analysen gemessen werden, die mit den Daten durchgeführt werden. Es ist uns daher ein langfristiges Anliegen, die Komplexität für die Nutzer weiterhin handhabbar zu gestalten. Ein nützliches Instrument (neben einer zugleich detaillierten und verständlichen Dokumentation) ist aus unserer Sicht die Veröffentlichung beispielhafter kommentierter Analysen in unseren „Quick Start“ Stata do-files, die auch Grundlage des vorliegenden Beitrags waren. Diesen Ansatz beabsichtigen wir in den nächsten Jahren fortzuführen.

VII. Literatur

Blossfeld, Hans-Peter, Katrin Golsch und Götz Rohwer, 2007: Event History Analysis with Stata. New York, London: Lawrence Erlbaum Associates.

Brüderl, Josef, Laura Castiglioni, Ulrich Krieger, Volker Ludwig, Klaus Pforr und Nina Schumann, 2010: pairfam Data Manual. Technical Report Mannheim Centre for European Social Research.

Brüderl, Josef und Stefani Scherer, 2004: Methoden zur Analyse von Sequenzdaten. Kölner Zeitschrift für Soziologie und Sozialpsychologie Sonderheft 44: 330-347.

- Brzinsky-Fay, Christian, Ulrich Kohler und Magdalena Luniak, 2006: Sequence analysis with Stata. The Stata Journal 6: 435-460.*
- Cleves, Mario A., William W. Gould, Roberto G. Gutierrez und Yulia U. Marchenko, 2008: An Introduction to Survival Analysis Using Stata. College Station, TX: Stata Press.*
- Heeringa, Steven G., Brady T. West und Patricia A. Berglund, 2010: Applied survey data analysis. Boca Raton, FL: Chapman & Hall/CRC.*
- Lynn, Peter, Nick Buck, Jonathan Burton, Annette Jäckle und Heather Laurie, 2005: A Review of Methodological Research Pertinent to Longitudinal Survey Design and Data Collection. ISER Working Paper 2005-29.*
- Scherer, Stefani, 2001: Early career patterns: A comparison of Great Britain and West Germany. European Sociological Review 17: 119-144.*
- Scherer, Stefani und Josef Brüderl, 2010: Sequenzdatenanalyse. S. 1031-1051 in: Wolf, Christof und Henning Best (Hg.), Handbuch der sozialwissenschaftlichen Datenanalyse VS Verlag für Sozialwissenschaften.*
- StataCorp., 2009: Stata: Release 11. Statistical Software. College Station, TX: StataCorp LP.*
- Suckow, Jana und Ulrich Schneekloth, 2009: Beziehungen und Familienleben in Deutschland (2008/2009) Welle 1. München: Infratest.*
- Walper, Sabine, Franziska Schmahl, Carolin Thönnissen, Markus Schaer und Eva-Verena Wendt, 2010: pairfam Skalenhandbuch Ankerpersonen-CAPI Welle 1. Ludwig-Maximilians-Universität München.*