

Forschungsdatenzentrum

der Bundesagentur für Arbeit
im Institut für Arbeitsmarkt-
und Berufsforschung

FDZ

FDZ-Methodenreport

04/2015

DE

Methodische Aspekte zu Arbeitsmarktdaten

PASS Quick Start File

Spellinformationen im Querschnitt

Arne Bethmann,
Maren Klawitter



Bundesagentur für Arbeit

PASS Quick Start File

Spellinformationen im Querschnitt

Arne Bethmann
Maren Klawitter

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

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

Inhaltsverzeichnis

Zusammenfassung	4
1 Hintergrund	5
1.1 Das Panel „Arbeitsmarkt und soziale Sicherung“	5
1.2 Datenstruktur	6
2 Beispiel	7
2.1 Datenaufbereitung	8
2.1.1 Vorbereiten des wellenspezifischen Interviewdatums der Personen-	
daten	8
2.1.2 Vorbereiten der Spell-Daten	9
2.1.3 Abgleich von Spell-Zeiträumen und Datumsangaben aus den Perso-	
neninterviews	10
2.2 Auswertung	12
2.2.1 Zusammenspielen der Analysedaten	12
2.2.2 Beispielanalysen	13
3 Variablenbeschreibung	17
3.1 Personendatensatz (PENDDAT)	17
3.1.1 Unveränderliche Personennummer	17
3.1.2 Indikator Befragungswelle	17
3.1.3 Interviewdatum Personeninterview – Jahr	17
3.1.4 Interviewdatum Personeninterview – Monat	18
3.1.5 Geschlecht des Befragten	18
3.2 Biographiedaten (bio_spells)	18
3.2.1 Spellnummer	18
3.2.2 Aktivitätstyp	19
3.2.3 Anfangsmonat	20
3.2.4 Anfangsjahr	20
3.2.5 Endmonat	21
3.2.6 Endjahr	21
3.2.7 Wie hat Befragter von Arbeitsstelle erfahren?	22
3.2.8 Tatsächliche Arbeitszeit	22
Literatur	23

Zusammenfassung

Der vorliegende FDZ-Methodenreport soll Nutzern des Panels „Arbeitsmarkt und soziale Sicherung“ (PASS) anhand eines Beispiels zeigen, wie Informationen aus den PASS-Spelldaten einer Welle zugeordnet und an die PASS-Querschnittdatensätze angespielt werden können.

Keywords: PASS, Panel, Spelldaten

1 Hintergrund

Die Forschungsdaten des PASS werden den Nutzern als Scientific Use File (SUF) zur Verfügung gestellt, das aus mehreren Datensätzen unterschiedlicher Formate besteht. Häufig erfordert die Analyse einer spezifischen Forschungsfrage eine Kombination verschiedener Teildatensätze, die für das Zusammenspielen im Vorfeld entsprechend aufbereitet werden müssen. Die komplexe Struktur der Daten stellt dabei regelmäßig eine Herausforderung für den betreffenden Forscher bzw. die betreffende Forscherin dar. In diesem Quick Start File (QSF) wird an einem Beispiel beschrieben, wie das Zusammenspielen von Spell-Informationen aus den Biographiedaten mit Querschnitts-Informationen zum Befragungszeitpunkt praktisch umgesetzt werden kann und welche Datenaufbereitungsschritte dazu notwendig sind.

Als Beispielfragestellung wird betrachtet, inwiefern das Geschlecht der befragten Person damit zusammenhängt, wie sie von ihrer aktuellen Arbeitsstelle erfahren hat. Dadurch kann gezeigt werden, wie die Verknüpfung von Querschnittsdaten aus den Personeninterviews mit Spelldaten aus der jeweiligen Erwerbsbiographie praktisch funktioniert. Die dabei verwendete Methodik bezieht sich zwar auf ein konkretes Beispiel, soll jedoch auch als Anhaltspunkt dienen, um andere Fragestellungen mit dem PASS zu bearbeiten, die eine Kombination von Spell- und Querschnittsdaten erfordern.

Im Folgenden wird zuerst ein kurzer Überblick über das PASS gegeben. Um die Problematik beim Zusammenspielen der verschiedenen Datensätze des SUF zu verdeutlichen, werden anschließend die verschiedenen Typen und Formate der Datensätze ausführlicher beschrieben. Darauf aufbauend wird die Kombination von personenspezifischen Querschnitts- und Spelldaten anhand eines kommentierten do-Files am angesprochenen Beispiel dargestellt.

1.1 Das Panel „Arbeitsmarkt und soziale Sicherung“

Das PASS ist eine seit 2006 jährlich stattfindende Haushaltsbefragung mit einer Überrepräsentation von Haushalten mit Arbeitslosengeld-II-Bezug (ALG II), welche vom Institut für Arbeitsmarkt- und Berufsforschung in Nürnberg (IAB) koordiniert und von infas Institut für angewandte Sozialwissenschaft GmbH durchgeführt wird. Aktuell stehen die Daten von sieben Erhebungswellen der wissenschaftlichen Forschung aufbereitet zur Verfügung. Vom Bundesministerium für Arbeit und Soziales (BMAS) in Auftrag gegeben, soll mit dem PASS u. a. Analysen im Rahmen der Gesetze für moderne Dienstleistungen am Arbeitsmarkt ermöglicht werden. Insbesondere die Dynamik der Grundsicherung, die Veränderung der wirtschaftlichen Lage und Bewältigungsprozesse der Haushalte aus dem Regelungsbereich des zweiten Sozialgesetzbuches (SGB II), sowie deren Kontakte zu den Trägern der Grundsicherung sollen untersucht werden. Somit stellen diese Daten ein wichtiges Instrument der Armut- und Arbeitsmarktforschung, sowie der Untersuchung der sozialen Lage von Haushalten und Personen mit ALG-II-Bezug dar (Berg u. a., 2014: 8ff.).

Im Rahmen des PASS werden sowohl Haushaltsinterviews mit dem Haushaltsvorstand, welcher die Person mit den umfassendsten Kenntnissen über die wirtschaftliche Situation des Haushaltes ist, als auch Personeninterviews mit jedem Haushaltmitglied über

15 Jahren durchgeführt. Personen ab 65 Jahren werden gebeten einen Kurzfragebogen (Seniorenfragebogen) zu beantworten. Informationen aus dem Haushaltsinterview betreffen beispielsweise den ALG-II-Bezug inklusive dessen Dauer, die Deprivation des Haushaltes sowie das Haushaltseinkommen, dessen Vermögen oder Schulden, die Wohnung und das Wohnumfeld oder Fragen zur Freizeitgestaltung sowie zur Betreuung von im Haushalt lebenden Kindern. Auf der Personenebene werden neben sozio-demographischen Merkmalen wie Alter, Geschlecht oder Bildung, u. a. detaillierte Informationen zu Erwerbstätigkeiten sowie zum (eventuellen) Leistungsbezug erhoben. Dabei werden beispielsweise Informationen über Arbeitszeit, Qualität der Beschäftigung, Erwerbsbiographie oder Arbeitssuche abgebildet (Trappmann u. a., 2013: 277).

1.2 Datenstruktur

Die Daten des PASS werden externen Nutzern aus wissenschaftlichen Institutionen als faktisch anonymisierter Scientific Use File (SUF) zur Verfügung gestellt, der aus mehreren Datensätzen unterschiedlichen Formats besteht.¹

Auf Grund der Befragungslogik können alle Datensätze des SUF entweder der Personen- oder der Haushaltsebene zugeordnet werden. Die Datensätze auf der Haushalts- oder Personenebene lassen sich wiederum in vier verschiedene Typen einteilen: Zum einen existieren Registerdatensätze, welche eine Liste von allen jemals befragten Haushalten, bzw. Personen enthalten. Daneben stellen die Gewichtungsdatsätze sowohl Haushalts-, als auch Personengewichte zur Verfügung. Damit sollen Verzerrungen auf Grund von Stichprobenziehung und Befragungsausfällen korrigiert werden. Zusätzlich erlaubt die Kalibrierung an statistische Eckwerte die Hochrechnung auf die Bevölkerung. Die Querschnittdatensätze enthalten alle personen- bzw. haushaltsbezogenen Informationen zum jeweiligen Erhebungszeitpunkt. Ergänzende Informationen zu berichteten Episoden von beispielsweise Arbeitslosigkeit oder Erwerbstätigkeit sind in Spelldatensätzen abgelegt. Hierbei werden für jede Episode sowohl ein Beginn- als auch ein Enddatum angegeben, die Episoden sind dabei vorerst nicht spezifischen Wellen zugeordnet.

Da in den Datensätzen des SUF die Informationen in verschiedener Art und Weise organisiert werden, weisen diese unterschiedliche Formate auf. Die Daten der Registerdatensätze sind im Wide-Format angeordnet. Für jede Befragungseinheit, also Personen oder Haushalte, werden die erhobenen Informationen in einer Zeile des Datensatzes dargestellt. Eine Zuordnung der Informationen zu dem jeweiligen Erhebungszeitpunkt kann anhand wellenspezifischer Variablen erfolgen. Demgegenüber sind die Querschnitts- und Gewichtungsdatsätze im Long-Format aufbereitet. Hierbei werden die erhobenen Informationen für jede Befragungseinheit wellenspezifisch dargestellt. Das bedeutet, dass jeweils ein Beobachtungszeitpunkt pro Befragungseinheit eine Zeile im Datensatz darstellt. Die Spell-Datensätze werden im Spell-Format zur Verfügung gestellt. In diesem Format repräsentiert jeweils eine Zeile eine berichtete Episode pro Befragungseinheit im Datensatz, wobei diese

¹ Im Bereich der Lehre an Hochschulen kann mit einem Campus File (CF) gearbeitet werden, der sich durch eine absolute Anonymisierung auszeichnet. Informationen zum Datenzugang sind hier zu finden: <http://fdz.iab.de/360/view.aspx>

sich auch über mehrere Erhebungswellen erstrecken kann (ausführlichere Informationen zur PASS-Datensatzstruktur sind in Bethmann/Fuchs/Wurdack, 2013: Kapitel 5 zu finden).

Sowohl auf der Haushalts- als auch auf der Personenebene werden im PASS spezifische Identifikationsnummern verwendet, welche eine Einheit über alle Wellen identifizieren. Anhand dieser Personen- oder Haushaltsnummern ist es möglich, Informationen aus verschiedenen Datensätzen einander zuzuordnen. Hierbei ist es abhängig vom Format des Datensatzes jedoch nicht immer möglich, auch den Zeitraum zu identifizieren, auf welchen sich eine Beobachtung bezieht. Datensätze im Long- oder Wide-Format weisen einen Wellenindikator auf, wodurch die Beobachtungen für jede Person einer Welle zugeordnet werden können. In den Registerdatensätzen kann der Befragungszeitpunkt über die wellenspezifischen Variablen bestimmt werden. Dagegen können Informationen aus Datensätzen im Spell-Format über das Beginn- und Enddatum zwar leicht dem Zeitraum zugeordnet werden, auf welchen sie sich beziehen, eine Zuordnung zu wellenspezifischen Querschnittsinformationen ist allerdings nicht so einfach möglich (Bethmann/Fuchs/Wurdack, 2013: Kapitel 9).

Technisch zeigt sich dieser Unterschied darin, dass Datensätze im Spell-Format keinen Wellenindikator beinhalten. Konzeptionell führt dies dazu, dass der Zeitpunkt, auf den sich die interessierende Querschnittsinformation bezieht, erst bestimmt werden muss – üblicherweise über das Datum des Interviews – um dann den Abgleich mit dem Gültigkeitszeitraum der Spell-Information durchzuführen. Das konkrete Vorgehen, um die berichteten Episoden aus den Spell-Datensätzen auf diesem Wege einer Welle zuzuordnen und mit den entsprechenden Querschnittsdaten zu verknüpfen, wird nachfolgend exemplarisch dargestellt.

2 Beispiel

Einige Forschungsfragen erfordern es, Informationen aus verschiedenen Datensätzen zu kombinieren. Ist beispielsweise von Interesse, wie das Geschlecht der Befragten damit zusammenhängt, wie diese von ihrer aktuellen Stelle erfahren haben, ist es erforderlich, die Querschnittsdaten auf der Personenebene mit Informationen aus den Spelldaten anzureichern. Das Geschlecht der Befragten ist in den Querschnittsdaten abgelegt. Informationen darüber, wie die aktuelle Stelle gefunden wurde, sind nur in den Spelldaten enthalten. Einzig die personenspezifische Identifikationsnummer stellt eine direkte Verknüpfung der Daten dar, da die Spelldaten nicht wie die Querschnittsdaten über einen Wellenindikator verfügen. Demzufolge muss zunächst eine Lösung gefunden werden, um die berichteten Episoden einem Erhebungszeitpunkt zuzuordnen. In einem nächsten Schritt können nun die relevanten Informationen aus den Spell- und Querschnittsdaten einander personen- und wellenspezifisch zugeordnet werden.

Anhand des folgenden kommentierten do-Files soll dieses Vorgehen dargestellt werden. Hierbei wird eine Möglichkeit aufgezeigt, wie berichtete Episoden der Erwerbstätigkeit einer Welle und so auch den wellenspezifischen Informationen aus den Querschnittsdaten auf der Personenebene zugeordnet werden können. Dieses do-File dient lediglich als Orientierungshilfe für ein mögliches Vorgehen. Je nach Fragestellung und Erkenntnisinteresse

kann auch eine andere oder eine auf diesem do-File basierende und an das individuelle Forschungsinteresse angepasste Lösung zum Zusammenspielen der Daten sinnvoll sein.

2.1 Datenaufbereitung

Zur besseren Reproduzierbarkeit und Dokumentation werden einige Stata-Einstellungen² gesetzt und ein log-File zum Protokollieren aller Stata-Ausgaben geöffnet.

```
version 12.1
clear all
set more off
capture log close
log using "$log\LOGFILE.log", replace
```

Außerdem werden einige globale Makros für die später verwendeten Pfadangaben definiert.

```
global orig "<Pfad mit PASS SUF Dateien>"
global data "<Pfad für erzeugte Datensätze>"
global log  "<Pfad für log-Datei>"
```

Es soll im Folgenden dargestellt werden, wie Erwerbstätigkeitsspell aus dem `bio_spell`-Datensatz mit dem Personendatensatz `PENDDAT` verknüpft und einer Welle zugeordnet werden können.

Um die beiden Datensätze zu kombinieren, müssen sie die gleiche Struktur aufweisen. Im vorliegenden Fall sollen ein Personen- und ein Wellenindikator vorliegen, anhand derer die Daten verknüpft werden können.

Der im Long-Format zur Verfügung gestellte Querschnittsdatsatz muss so aufbereitet werden, dass eine Beobachtung pro Person und Welle durch eine Zeile des Datensatzes dargestellt wird. In den Spelldaten ist kein Wellenindikator vorhanden. Es muss also eine Möglichkeit gefunden werden, die berichteten Episoden einer Welle zuzuordnen. Möglich wird diese Zuordnung durch die Angaben zu Beginn- und Enddatum eines Spells, da in den Querschnittsdaten das genaue Interviewdatum für jede Person wellenspezifisch erfasst wird.

2.1.1 Vorbereiten des wellenspezifischen Interviewdatums der Personendaten

Zunächst wird der Personendatsatz vorbereitet. Der erste Schritt besteht darin, die Daten zu laden. Dabei werden die Variablen bereits auf die notwendigen Variablen – `pnr welle pintjahr pintmon` – beschränkt (`use <notwendige Variablen> using <Datensatz>`).

² Für der Erstellung der Beispiele wurde Stata 12.1 verwendet. Sollten Sie über eine andere Stata-Version verfügen, müssen Sie diese Einstellungen evtl. anpassen.

Dann wird eine neue Variable erstellt, welche das individuelle Interviewdatum für die jeweilige Welle abbildet. Anhand dieses Datums kann später festgestellt werden, ob zu dem jeweiligen Interviewzeitpunkt ein Erwerbsspell angedauert hat. Der Monat und das Jahr des Personeninterviews in der jeweiligen Erhebungswelle werden mit der Stata-Funktion `ym()` zusammengefasst. Da in den Spelldaten nur monatsgenaue Beginn- und Enddaten der Erwerbsepisoden erfasst werden, kann die tagesgenaue Angabe des Interviewdatums ignoriert werden. Nach der Generierung werden die ursprünglichen Monats- und Jahresangaben nicht mehr benötigt und können aus dem Datensatz entfernt werden.

```
use pnr welle pintjahr pintmon using "$orig\PENDDAT.dta", clear
gen intdat = ym(pintjahr, pintmon)
drop pintjahr pintmon
```

Für den Abgleich mit den Spelldaten werden die Datumsangaben zum Personeninterview ins Wide-Format gebracht. Das bedeutet, dass mit Hilfe des `reshape`-Befehls die gerade generierten Angaben zum Interviewdatum (`intdat`) aus den wellenspezifischen Zeilen (Option: `j(welle)`) für jede Person (Option: `i(pnr)`) in wellenspezifische Spalten (lies: Variablen) überführt werden.

```
reshape wide intdat, i(pnr) j(welle)
```

Baei werden mehrere, wellenspezifische Zeilen je Person mit jeweils einer Datumsangabe zu einer einzigen Zeile pro Person mit mehreren, wellenspezifischen Datumsangaben zusammengefasst:

Long-Format				Wide-Format				
pnr	welle	intdat		pnr	intdat1	intdat2	intdat3	intdat...
123	1	569						
123	2	579						
123	3	590	→	123	569	579	590	...
231	2	582						
231	3	591	→	231	.	582	591	...
...

Anschließend wird der soeben erzeugte Datensatz für die spätere Nutzung als `intdatwelle.dta` zwischengespeichert.

```
save "$data\intdatwelle.dta", replace
```

2.1.2 Vorbereiten der Spell-Daten

Im nächsten Schritt wird der Spell-Datensatz geöffnet und vorbereitet. Auch hier werden zunächst nur die notwendigen Variablen – `pnr spellnr bmonat bjahr emonat ejahr`

spelltyp – geladen. Die spellnr ist ein Zähler, der es ermöglicht, innerhalb einer Person einen Spell eindeutig zu identifizieren und später benötigt wird, um die Spells mit Querschnittsinformationen aus den Personendaten zusammenszuspielen. Für das Beispiel werden nur Erwerbstätigkeitsepisoden benötigt, weshalb auch nur dieser Spelltyp behalten wird (spelltyp == 1).

```
use pnr spellnr bmonat bjahr emonat ejahr spelltyp \\\  
    using "$orig\bio_spells.dta"  
keep if spelltyp == 1
```

Fälle mit fehlenden Angaben zu den Beginn- oder Endzeitpunkten eines Spells werden gelöscht, da sie nicht ohne weiteres einer Welle zugeordnet werden können.³

```
drop if bmonat < 0 | bjahr < 0 | emonat < 0 | ejahr < 0
```

Anschließend wird – analog zum Interviewdatum in den Personendaten – je eine Variable für das Beginn- und Enddatum eines Spells gebildet. Die ursprünglichen Monats- und Jahresangaben sowie der Spelltyp werden nun nicht mehr benötigt und daher gelöscht.

```
gen begdat = ym(bjahr, bmonat)  
gen enddat = ym(ejahr, emonat)  
drop bmonat bjahr emonat ejahr spelltyp
```

2.1.3 Abgleich von Spell-Zeiträumen und Datumsangaben aus den Personeninterviews

Anschließend werden den Spelldaten die im ersten Schritt generierten Interviewdaten zugespielt. Da beide Datensätze über den Personenindikator pnr verfügen, kann dieser für den merge-Befehl verwendet werden. Im Rahmen der PASS-Befragung kann eine Person bereits mehrere Erwerbstätigkeitsepisoden berichtet haben, dies wird beim Zusammenspielen über das Merge-Verfahren m:1 berücksichtigt. Für die weitere Datenaufbereitung sind Personen ohne berichtete Erwerbstätigkeitsspells nicht verwendbar, weshalb sie beim Zusammenspielen über die Option keep(1 3) direkt ausgeschlossen werden.

```
merge m:1 pnr using "$data\intdatwelle.dta", nogen keep(1 3)
```

Fälle mit der gleichen pnr in Spell- und Personendatensatz werden zusammengespielt, indem die wellenspezifischen Datumsangaben aus dem Personendatensatz als neue Variablen an die jeweilige(n) Zeile(n) im Spell-Datensatz angehängt werden. Kommt eine pnr im Spell-Datensatz mehrfach vor, erhält jede Zeile die gleichen Datumsangaben aus dem Personendatensatz:

³ Auf mögliche Imputationsverfahren kann aus Platzgründen an dieser Stelle nicht eingegangen werden.

Spell-Daten				Personendaten				
pnr	spellnr	begdat	enddat		pnr	intdat1	intdat2	intdat...
324	1	569	593	←	324	564	577	...
432	1	579	587	←	432	598	609	...
432	2	595	612					
...

merge m:1 pnr

pnr	spellnr	begdat	enddat	intdat1	intdat2	intdat...
324	1	569	593	564	577	...
432	1	579	587	598	609	...
432	2	595	612	598	609	...

Nun werden für jeden Spell sieben Indikatoren `et_w1` bis `et_w7` gebildet (einer pro Welle), die anzeigen, ob die Erwerbstätigkeit zum Interviewzeitpunkt der jeweiligen Welle angedauert hat oder nicht. Dazu wird für jedes Interviewdatum geprüft, ob es zwischen Beginn- und Enddatum des Spells liegt. Wenn das so ist, hat der Spell zum Interviewzeitpunkt ange-dauert und die befragte Person war erwerbstätig. Sollte das Interviewdatum auf „fehlender Wert“ (in Stata: „.“) stehen, bedeutet das, dass die jeweilige Person in der Welle kein Personeninterview gegeben hat. Der Indikator, ob der Spell andauert, wird dementsprechend auch auf „.“ gesetzt. Nach der Generierung werden weder die Variablen mit Beginn- und Enddatum des Spells noch die Angaben zum Interviewdatum benötigt und daher gelöscht.

```
gen et_w1 = intdat1 >= begdat & intdat1 <= enddat if intdat1 != .
gen et_w2 = intdat2 >= begdat & intdat2 <= enddat if intdat2 != .
gen et_w3 = intdat3 >= begdat & intdat3 <= enddat if intdat3 != .
gen et_w4 = intdat4 >= begdat & intdat4 <= enddat if intdat4 != .
gen et_w5 = intdat5 >= begdat & intdat5 <= enddat if intdat5 != .
gen et_w6 = intdat6 >= begdat & intdat6 <= enddat if intdat6 != .
gen et_w7 = intdat7 >= begdat & intdat7 <= enddat if intdat7 != .
drop begdat enddat intdat*
```

Um mit den Daten bequemer weiterarbeiten zu können, werden sie nun wieder ins Long-Format gebracht. Der `reshape`-Befehl benötigt dazu einen eindeutigen Indikator, der bisher allerdings noch nicht vorhanden ist. Die `pnr` eignet sich in diesem Fall nicht, da der Datensatz nicht personen- sondern spell-spezifisch umgeformt werden soll. Die Daten sollen angeben, ob der jeweilige Spell zum Befragungszeitpunkt in einer bestimmten Welle angedauert hat. Ein neuer eindeutiger Spellindikator lässt sich erzeugen, indem man die `pnr` mit der `spellnr` kombiniert. Anschließend wird der `reshape`-Befehl in der `long`-Richtung ausgeführt. Jede Zeile – und damit jeder Spell – ist dabei eindeutig über die `id` definiert.

```
gen double id = (pnr * 1000) + spellnr
reshape long et_w, i(id) j(welle)
```

Für jeden Spell werden so viele Zeilen erzeugt, wie Befragungswellen im Datensatz verfügbar sind, aktuell also sieben. Die wellenspezifischen Indikatoren, ob der Spell zum jeweiligen Interviewdatum andauert, werden dabei aus den Spalten in die jeweiligen Zeilen kopiert:

Wide-Format

id	pnr	spellnr	et_w1	et_w2	et_w...
345004	345	4	0	1	...
435002	435	2	.	1	...
435006	435	6	.	0	...
...

↓

Long-Format

id	pnr	spellnr	et_w	welle
345004	345	4	0	1
345004	345	4	1	2
435002	345	2	.	1
435002	345	2	1	2
435006	345	6	.	1
435006	345	6	0	2
...

Für die weiteren Analysen sind nur Spells interessant, die zu einem Interviewdatum andauern haben (`et_w == 1`). Daher werden auch nur die entsprechenden Zeilen im Datensatz behalten. Für das Zusammenspielen von Querschnittsinformationen aus dem Personendatensatz und Spell-Informationen aus der Erwerbshistorie werden nur die Merkmale `pnr` `welle` `spellnr` benötigt. Der Datensatz wird auf diese Variablen reduziert und abschließend unter `etspellwelle.dta` gespeichert.

```
keep if et_w == 1
keep pnr welle spellnr
save "$data\etspellwelle.dta", replace
```

2.2 Auswertung

Der so erzeugte Datensatz bildet eine Schlüsseldatei, um den wellenspezifischen Querschnittsinformationen aus den Personeninterviews Spell-Informationen zu den jeweils andauernden Erwerbstätigkeitsepisoden zuzuspielen.

2.2.1 Zusammenspielen der Analysedaten

Um die Daten für die Beispielanalysen zusammenzuspielen, werden nun zuerst die Variablen `pnr` `welle` `zpsex` aus den Querschnittsinformationen geladen. Von inhaltlichem Interesse

ist dabei nur die Variable `zpsex`, die das Geschlecht der befragten Person angibt.

```
use pnr welle zpsex using "$orig\PENDDAT.dta", clear
```

In einem Zwischenschritt werden nun die Informationen zu den andauernden Erwerbstätigkeit über `pnr` und `welle` zugespielt. Da es möglich ist, dass Personen mehr als eine Erwerbstätigkeit gleichzeitig ausüben, wird über die `merge`-Variante `1:m` zugelassen, dass jeder Zeile aus den Personendaten mehr als eine Zeile aus der Schlüsseldatei zugespielt wird.

```
merge 1:m pnr welle using "$data\etspellwelle.dta", nogen
```

Nun ist es möglich die Informationen aus den Erwerbstätigkeitsspiels zuzuspielen, da sie über die `pnr` in Kombination mit der eben zugespielten `spellnr` eindeutig identifiziert werden können. Erwerbstätigkeiten können durchaus über mehrere Wellen andauern, weshalb dann Informationen aus nur einem Spell (lies: Zeile in den Spelldaten) mehreren Zeilen in den Personendaten zugespielt werden müssen. Dies wird hier über die `merge`-Variante `m:1` ermöglicht. Dabei werden beim Zusammenspielen alle Spells, die nicht eine Erwerbstätigkeit beschreiben oder aber zumindest zu keinem Interviewzeitpunkt andauern, ignoriert (`keep(1 3)`). Es sollen nur zwei Variablen aus dem Spell-Datensatz übernommen werden: `ET2401`, die angibt, über welchen Weg die aktuelle Arbeitsstelle gefunden wurde und `az2`, welche die tatsächliche, momentane Arbeitszeit angibt und benötigt wird, um die Haupterwerbstätigkeit zu identifizieren.

```
merge m:1 pnr spellnr using "$orig\bio_spells.dta", nogen ///  
keep(1 3) keepus(ET2401 az2)
```

2.2.2 Beispielanalysen

Bei parallel ausgeübten Erwerbstätigkeiten soll im Folgenden immer nur die Erwerbstätigkeit mit dem größten Arbeitszeitumfang berücksichtigt werden. Dazu werden die Zeilen innerhalb jeder Person und Welle nach dem Arbeitszeitumfang absteigend sortiert. Anschließend werden alle Zeilen bis auf die oberste sowie die Arbeitszeitvariable gelöscht.

```
gsort pnr welle -az2  
by pnr welle: keep if _n == 1  
drop az2
```

Der nun fertige Analysedatensatz wird zwischengespeichert.

```
save "$data\datensatz.dta", replace
```

Um einen Überblick über die Daten zu bekommen, werden Häufigkeitsauszählungen für die Analysevariablen `zpsex` und `ET2401` ausgegeben. Bei der zweiten Variable fällt auf, dass es sehr viele systembedingt fehlende Werte („.“) gibt. Das liegt daran, dass im Datensatz noch alle geführten Personeninterviews enthalten sind, unabhängig davon, ob eine andauernde Erwerbstätigkeit zugespielt werden konnte oder nicht. Für die Geschlechtsangabe und den Weg, wie die aktuelle Arbeitsstelle gefunden wurde, wird jeweils ein neue Variable angelegt, die alle fehlenden Werte auf systembedingte Missings setzt. Diese Fälle werden so in den weiteren Analysen nicht mehr berücksichtigt. Die `ET2401` wird außerdem auf wenige Kategorien reduziert, um das Beispiel zu vereinfachen.

```
tab zpsex
gen frau = zpsex == 2 if zpsex >= 0
tab ET2401, m
recode ET2401 ( 1 = 1 "Zeitung") ///
              ( 3 = 2 "Internet") ///
              ( 4 = 3 "Bekannte") ///
              ( 7 = 4 "Nachfr. bei Betrieb") ///
              (12 = 5 "vorher. Arbeitgeber") ///
              (2 5 6 8 9 10 11 = 6 "Sonstiges") ///
              (else = .) ///
              , gen(eterf)
```

Der Zusammenhang von Geschlecht und dem Weg, auf dem die aktuelle Arbeitsstelle gefunden wurde, kann nun anhand von Kreuztabellen – exemplarisch für die Wellen 6 und 7 – analysiert werden. Die Auszählungen sind in Tabelle 1 abgebildet und enthalten zusätzlich einige statistische Kennzahlen.⁴

```
tab eterf frau if welle == 6, col chi2 V
tab eterf frau if welle == 7, col chi2 V
```

Im Folgenden werden die Zusammenhänge grafisch dargestellt. Dazu werden gruppierte Balkendiagramme verwendet. Um den `graph hbar`-Befehl von Stata zu verwenden müssen im ersten Schritt Indikatorvariablen für alle Ausprägungen der Variable `eterf` generiert werden (`erf1` bis `erf6`). Anschließend wird ein gelabeltes und nach Welle und Geschlecht gruppiertes, horizontales Balkendiagramm erzeugt. Im letzten Schritt wird dieses als PDF-Datei gespeichert. Das Ergebnis ist in Abbildung 1 zu sehen.

```
tab eterf, gen(erf)
graph hbar erf1-erf6 if welle == 6 | welle == 7, ///
  over(frau, relabel(1 "Männer" 2 "Frauen")) ///
```

⁴ Bitte beachten Sie, dass es sich hier um ungewichtete Analysen handelt. Für eine seriöse inhaltliche Interpretation ist es dringend geraten die mitgelieferten Gewichtungsfaktoren zu verwenden. Nähere Informationen für die Gewichtung in PASS sind bei Bethmann/Fuchs/Wurdack (2013: Kapitel 12) zu finden.

Tabelle 1: Beispiel – Wie von Stelle erfahren? Nach Geschlecht und Welle

	Welle 6 (2012)			Welle 7 (2013)		
	Männer	Frauen	Gesamt	Männer	Frauen	Gesamt
Zeitung	8,2 % 100	12,9 % 154	10,5 % 254	8,4 % 111	12,4 % 173	10,4 % 284
Internet	8,6 % 105	8,0 % 95	8,3 % 200	7,8 % 103	7,8 % 109	7,8 % 212
Bekannte	27,9 % 340	23,9 % 284	25,9 % 624	25,6 % 340	23,6 % 329	24,6 % 669
Nachfrage bei Betrieb	18,1 % 221	18,8 % 224	18,5 % 445	19,4 % 258	20,8 % 290	20,1 % 548
vorheriger Arbeitgeber	11,0 % 134	13,1 % 156	12,0 % 290	11,6 % 154	13,7 % 191	12,7 % 345
Sonstiges	26,3 % 321	23,3 % 278	24,8 % 599	27,3 % 362	21,8 % 304	24,5 % 666
Gesamt	100,0 % 1 221	100,0 % 1 191	100,0 % 2 412	100,0 % 1 328	100,0 % 1 396	100,0 % 2 724
	$\chi^2 = 21,4$ (df = 5), Pr = 0,001 Cramers V = 0,09			$\chi^2 = 23,1$ (df = 5), Pr = 0,000 Cramers V = 0,09		

Spaltenprozentage und Absolutwerte

```

over(welle, relabel(1 "2012" 2 "2013")) ///
yvaroptions(relabel(1 "Zeitung" ///
                    2 "Internet" ///
                    3 "Bekannte" ///
                    4 "Nachfr. bei Betrieb" ///
                    5 "vorher. Arbeitgeber" ///
                    6 "Sonstiges")) ///

title("Wie haben Sie von Ihrer Arbeitsstelle erfahren?") ///
graphregion(color(white)) ytitle("Anteil in %") ///
ylabel(0 .05 "5" .1 "10" .15 "15" .2 "20" .25 "25" .3 "30")
graph export "$data\eterf.pdf", replace

```

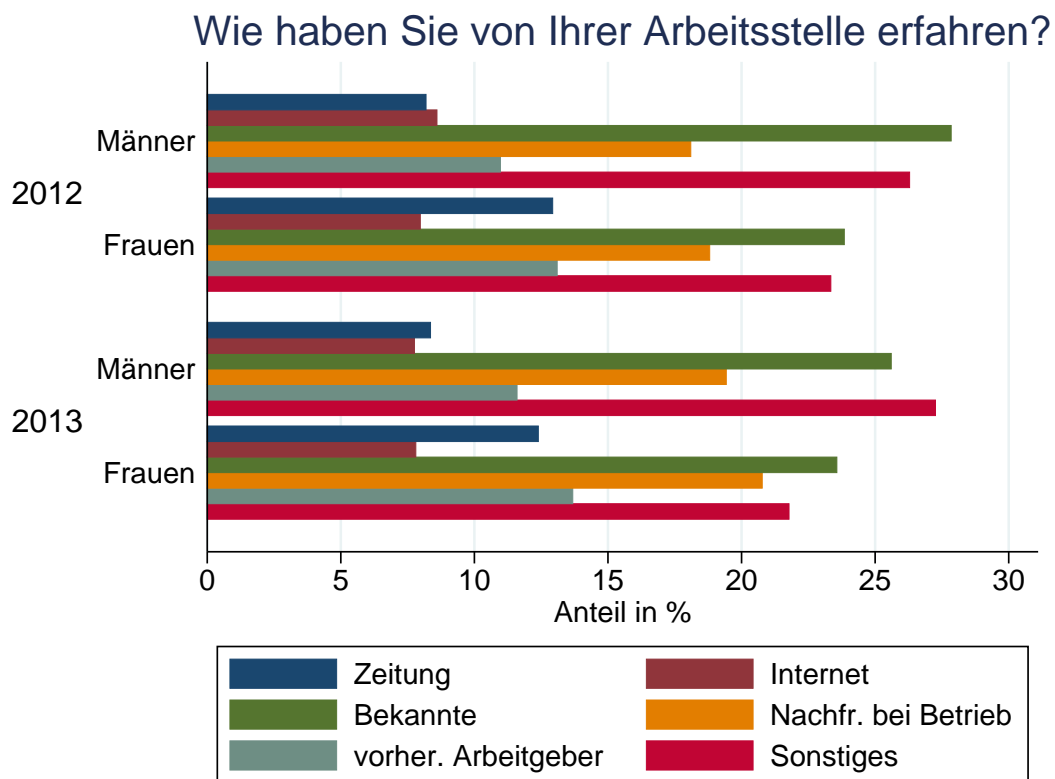
Zum Abschluss des do-Files werden dann noch log-File und Stata geschlossen.

```

log close
exit, clear

```

Abbildung 1: Beispiel – Wie von Stelle erfahren? nach Geschlecht und Welle



3 Variablenbeschreibung

3.1 Personendatensatz (PENDDAT)

3.1.1 Unveränderliche Personennummer

Variablenlabel	Unveraenderliche Personennummer
Variablenname	pnr
Detailbeschreibung	Dieses Merkmal bildet die unveränderliche Personennr. über alle Erhebungszeitpunkte ab.
Kategorien	<i>Identifikator</i>

3.1.2 Indikator Befragungswelle

Variablenlabel	Indikator Befragungswelle
Variablenname	welle
Detailbeschreibung	Dieses Merkmal bildet die Erhebungswelle ab.
Kategorien	1 – Welle 1 (2006/2007) 2 – Welle 2 (2007/2008) 3 – Welle 3 (2008/2009) 4 – Welle 4 (2010) 5 – Welle 5 (2011) 6 – Welle 6 (2012) 7 – Welle 7 (2013)

3.1.3 Interviewdatum Personeninterview – Jahr

Variablenlabel	Interviewdatum Personeninterview – Jahr, generiert
Variablenname	pintjahr
Detailbeschreibung	Dieses Merkmal bildet das Jahr des jeweiligen Personeninterviews ab.
Kategorien	2006 – 2013

3.1.4 Interviewdatum Personeninterview – Monat

Variablenlabel	Interviewdatum Personeninterview – Monat, generiert
Variablenname	pintmon
Detailbeschreibung	Dieses Merkmal bildet den Monat des jeweiligen Personeninterviews ab.
Kategorien	1 – Januar 2 – Februar 3 – März 4 – April 5 – Mai 6 – Juni 7 – Juli 8 – August 9 – September 10 – Oktober 11 – November 12 – Dezember

3.1.5 Geschlecht des Befragten

Variablenlabel	Steuervariable: Geschlecht des Befragten (aus HHgrid)
Variablenname	zpsex
Detailbeschreibung	Diese Variable bildet das Geschlecht des Befragten ab.
Kategorien	1 – männlich 2 – weiblich Missing Values: -2 – Keine Angabe -4 – Frage irrtümlich nicht gestellt

3.2 Biographiedaten (bio_spells)

3.2.1 Spellnummer

Variablenlabel	Alle Aktivitaeten: Spellnummer
Variablenname	spellnr
Detailbeschreibung	Nummer der jeweils berichteten Episode.
Kategorien	<i>Zähler</i>

3.2.2 Aktivitätstyp

Variablenlabel	Alle Aktivitaeten: Aktivitaetstyp, generiert
Variablenname	spelltyp
Detailbeschreibung	Aktivitätstyp (erwerbstätig, arbeitslos, etc.) der Episode.
Kategorien	<p>1 – Erwerbstätig mit Einkommen > 400 Euro/Monat</p> <p>2 – Arbeitslos gemeldet / Massnahmenteilnahme</p> <p>3 – Arbeitslos, nicht gemeldet (ab Welle 4 aus offenen Angaben)</p> <p>4 – Ausbildung / Lehre / Studium</p> <p>5 – Wehr- / Zivildienst / etc.</p> <p>6 – Hausfrau / Hausmann</p> <p>7 – Mutterschutz / Erziehungsurlaub / Elternzeit</p> <p>8 – Rentner / Pensionär / Vorruhestand</p> <p>9 – Etwas anderes gemacht</p> <p>10 – Schüler (bis Welle 3 aus offenen Angaben)</p> <p>11 – Hauptstatus unklar (aus offenen Angaben)</p> <p>12 – Krank / berufsunfähig / erwerbsunfähig / behindert (aus offenen Angaben)</p> <p>13 – Selbständig / mithelfende Familienangehörige (aus offenen Angaben)</p> <p>Missing Values:</p> <p>-1 – Weiß nicht</p> <p>-2 – Keine Angabe</p>

3.2.3 Anfangsmonat

Variablenlabel	Alle Aktivitaeten: Anfangsmonat, generiert
Variablenname	bmonat
Detailbeschreibung	Beginnmonat der Episode, in der für die Generierung Jahreszeitangaben in korrekte Monatsangaben umgerechnet wurden.
Kategorien	1 – Januar 2 – Februar 3 – März 4 – April 5 – Mai 6 – Juni 7 – Juli 8 – August 9 – September 10 – Oktober 11 – November 12 – Dezember Missing Values: -1 – Weiß nicht -2 – Keine Angabe -3 – Trifft nicht zu (Filter) -5 – Spellbeginn auf Grund der Befragungslogik nicht erhoben -8 – Unplausibler Wert

3.2.4 Anfangsjahr

Variablenlabel	Alle Aktivitaeten: Anfangsjahr, generiert
Variablenname	bjahr
Detailbeschreibung	Beginnjahr der Episode.
Kategorien	1913 – 2013 Missing Values: -1 – Weiß nicht -2 – Keine Angabe -3 – Trifft nicht zu (Filter) -5 – Spellbeginn auf Grund der Befragungslogik nicht erhoben -8 – Unplausibler Wert

3.2.5 Endmonat

Variablenlabel	Alle Aktivitaeten: Endmonat, generiert
Variablenname	emonat
Detailbeschreibung	Endmonat der Episode, in der für die Generierung Jahreszeitangaben in korrekte Monatsangaben umgerechnet wurden.
Kategorien	1 – Januar 2 – Februar 3 – März 4 – April 5 – Mai 6 – Juni 7 – Juli 8 – August 9 – September 10 – Oktober 11 – November 12 – Dezember Missing Values: -1 – Weiß nicht -2 – Keine Angabe -3 – Trifft nicht zu (Filter) -4 – Frage irrtümlich nicht gestellt -8 – Unplausibler Wert

3.2.6 Endjahr

Variablenlabel	Alle Aktivitaeten: Endjahr, generiert
Variablenname	ejahr
Detailbeschreibung	Endjahr der Episode.
Kategorien	2005 – 2013 Missing Values: -1 – Weiß nicht -2 – Keine Angabe -3 – Trifft nicht zu (Filter) -4 – Frage irrtümlich nicht gestellt -8 – Unplausibler Wert

3.2.7 Wie hat Befragter von Arbeitsstelle erfahren?

Variablenlabel	Wie von der Stelle erfahren?, +offen
Variablenname	ET2401
Detailbeschreibung	Information, wie der Befragte von der betreffenden Arbeitsstelle erfahren hat. Offene Angaben wurden wo möglich geschlossenen Kategorien zugeordnet.
Kategorien	<p>1 – Aus einer Stellenanzeige in der Zeitung? 2 – Aus der Online-Stellenbörse der Agentur für Arbeit? 3 – Aus anderen Internetquellen? 4 – Von Bekannten oder Verwandten? 5 – Von einem Vermittler der Agentur für Arbeit? 6 – Von einem privaten Vermittler? 7 – Durch eine eigene Nachfrage in einem Betrieb? 8 – Sonstiges 9 – Nichts davon 10 – Übernahme aus einer Zeitarbeitstätigkeit 11 – Zielperson hat sich selbständig gemacht 12 – Durch den (vorherigen / aktuellen) Arbeitgeber</p> <p>Missing Values: -1 – Weiß nicht -2 – Keine Angabe -3 – Trifft nicht zu (Filter) -6 – Erwerbstätigkeit aus offener Angabe; Lückenmodul -9 – Item in Welle nicht erhoben</p>

3.2.8 Tatsächliche Arbeitszeit

Variablenlabel	Tatsächliche Arbeitszeit inkl. Angaben bei unregelmässiger Arbeitszeit, generiert
Variablenname	az2
Detailbeschreibung	Tatsächliche wöchentliche Arbeitszeit, bei der auch Antworten berücksichtigt wurden, die aus der Nachfrage bei unregelmässiger Arbeitszeit stammen.
Kategorien	<p>0 – 96</p> <p>Missing Values: -1 – Weiß nicht -2 – Keine Angabe -3 – Trifft nicht zu (Filter) -5 – Nicht codierbar</p>

Literatur

Berg, Marco; Cramer, Ralph; Dickmann, Christian; Gilberg, Reiner; Jesske, Birgit; Kleudgen, Martin; Bethmann, Arne; Fuchs, Benjamin; Huber, Martina; Trappmann, Mark (2014): Codebuch und Dokumentation des „Panel Arbeitsmarkt und soziale Sicherung“ (PASS). Datenreport Welle 7. FDZ-Datenreport 02/2014, Institut für Arbeitsmarkt- und Berufsforschung, Nürnberg.

Bethmann, Arne; Fuchs, Benjamin; Wurdack, Anja (2013): User Guide „Panel Study Labour Market and Social Security“ (PASS). Wave 6. FDZ-Datenreport 07/2013, Institut für Arbeitsmarkt- und Berufsforschung, Nürnberg.

Trappmann, Mark; Beste, Jonas; Bethmann, Arne; Müller, Gerrit (2013): The PASS Panel Survey After Six Waves. In: Journal for Labour Market Research, Bd. 45, Nr. 4, S. 275–281.

Impressum

FDZ-Methodenreport 4/2015

Herausgeber

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

Redaktion

Stefan Bender, Dagmar Theune

Technische Herstellung

Dagmar Theune

Rechte

Nachdruck - auch auszugsweise - nur mit
Genehmigung des FDZ gestattet

Bezugsmöglichkeit

http://doku.iab.de/fdz/reporte/2015/MR_04-15.pdf

Internet

<http://fdz.iab.de/>

Rückfragen zum Inhalt an:

Dr. Arne Bethmann
Institut für Arbeitsmarkt- und
Berufsforschung (IAB)
Regensburger Str. 104
90478 Nürnberg
Telefon: 0911 / 179-2307
E-Mail: Arne.Bethmann@iab.de