IABForschungsbericht

Ergebnisse aus der Projektarbeit des Instituts für Arbeitsmarkt- und Berufsforschung

Nr. 18/2006

Der Einsatz von Missing Data Techniken in der Arbeitsmarktforschung des IAB

Susanne Rässler

Bundesagentur für Arbeit

Der Einsatz von Missing Data Techniken in der Arbeitsmarktforschung des IAB

Susanne Rässler (IAB)

Mit der Publikation von Forschungsberichten will das IAB der Fachöffentlichkeit Einblick in seine laufenden Arbeiten geben. Die Berichte sollen aber auch den Forscherinnen und Forschern einen unkomplizierten und raschen Zugang zum Markt verschaffen. Vor allem längere Zwischen- aber auch Endberichte aus der empirischen Projektarbeit bilden die Basis der Reihe, die den bisherigen „IAB-Werkstattbericht“ ablöst.

Der Einsatz von Missing Data Techniken in der Arbeitsmarktforschung des IAB

Susanne Rässler *

Zusammenfassung

Dieser Beitrag gibt einen Überblick über aktuelle Forschungsprojekte des Instituts für Arbeitsmarkt- und Berufsforschung (IAB), die sich alle auf unterschiedliche Weise mit der Behebung von Problemen mit Missing Data befassen. Hierzu gehören einerseits Projekte, die Stichproben verwenden wie das IAB-Betriebspanel, die IAB-Beschäftigtenstichprobe oder die Erhebung der offenen Stellen, die entweder wie alle Stichprobenerhebungen durch Antwortausfälle beeinträchtigt sind oder aber zensierte Beobachtungen aufweisen. Auf der anderen Seite zählen hierzu Forschungsprojekte, die versuchen, unterschiedliche Datengrundlagen vereinbar zu machen, die auf verschiedenen regionalen oder klassifikatorischen Abgrenzungen beruhen. Außerdem wird das Projekt TrEffeR des IAB und der Bundesagentur für Arbeit (BA) vorgestellt, das der oft ausgesprochenen Forderung gerecht wird, die Maßnahmen der aktiven Arbeitsmarktpolitik auf breiter Basis zu evaluieren und diese Erkenntnisse für das operative Geschäft der BA nutzbar macht.

Keywords: Multiple Imputation, Missing by Design, Markov Chain Monte Carlo Methoden, Evaluation aktiver Arbeitsmarktpolitik, Arbeitslosigkeit, Prozessdaten, Propensity-Score-Matching.

JEL codes: C15, J68, J64, C11, C63, C14, J41

*Vortrag für die Statistische Woche 2005, die Jahrestagung der Deutschen Statistischen Gesellschaft. Ich danke den Mitarbeitern meiner beiden Abteilungen IAB-KEM und BA-PP4 sowie meinen Co-Autoren und Forschungspartnern der jeweiligen Projekte. Mein ganz besonderer Dank geht an Katrin Schmidt, die mich bei der Erstellung des Artikels wesentlich unterstützt hat sowie an Donald B. Rubin für zahlreiche hilfreiche und anregende Diskussionen sowie die Unterstützung bei der Entwicklung des Projektes TrEffeR. Einen herzlichen Dank darf ich auch den vielen Kolleginnen und Kollegen im IAB und der Zentrale der Bundesagentur für Arbeit sagen, ohne deren Aufmunterung, Zustimmung und Mithilfe diese vielen neuen Projekte kaum machbar wären.

Inhaltsverzeichnis

1 Einleitung

Die Erhebung statistischer Informationen erfolgt - aus Gründen der Praktikabilität oder aus Kostengründen - häufig auf Stichprobenbasis. Was unter Kostenaspekten sinnvoll ist, führt jedoch hinsichtlich der Datenqualität häufig zu Problemen, da Stichprobendaten oft Mängel aufweisen. Zu diesen gehören unter anderem Coverage-, Stichproben- oder Interviewerfehler sowie Probleme, die sich aus dem Stichprobendesign ergeben. Ein zentrales Problem bei Stichprobenerhebungen ist jedoch das der fehlenden Daten, der *missing data*. Hierbei wird unterschieden in die zwei Problemkreise *unit nonresponse* und *item nonresponse*. Ersteres bezeichnet das Fehlen von ganzen Datenreihen, wenn für einen Merkmalsträger keine Beobachtungen der Erhebungsvariablen vorliegen. Dies ist beispielsweise dann der Fall, wenn eine für eine Befragung ausgewählte Person die Teilnahme komplett verweigert oder trotz mehrfacher Versuche nicht anzutreffen ist. Zweiteres bezeichnet die Lückenhaftigkeit von Daten, wenn zu einer Untersuchungseinheit zwar Beobachtungen vorliegen, diese jedoch unvollständig sind. Dieser Fall tritt dann auf, wenn beispielsweise Fragen nach dem Einkommen oder den Vermögensverhältnissen nicht, andere Fragen jedoch beantwortet werden.

Die Diskussion dieser Problematik in der Literatur erfolgt bereits seit mehreren Jahrzehnten (vgl. beispielsweise Hartley und Hocking, 1971; Rubin, 1972, 1974, 1976; Little, 1976; Kalton, 1983; und Madow *et al.*, 1983), erlangt aber vor allem in jüngerer Vergangenheit zunehmende Bedeutung. Ein Grund hierfür ist sicherlich die Weiterentwicklung der technischen Möglichkeiten mit lückenhaften Datensätzen umzugehen und die dadurch ansteigende Zahl praktischer Applikationen der Verfahren.

Doch auch prozessproduzierte Daten können Lücken oder fehlende Wertebereiche aufweisen. Häufig empfiehlt es sich daher, statistische Fragestellungen auch oder insbesondere unter einer Missing Data Perspektive zu betrachten, wie dieser Beitrag zeigen soll.

Die vorgestellten Projekte stellen einige unterschiedliche Forschungsarbeiten des Instituts für Arbeitsmarkt- und Berufsforschung (IAB) der Bundesagentur für Arbeit (BA) vor. Sie haben gemeinsam, dass alle auf Daten basieren, die aufgrund ihres Erhebungsmodus' Missing Data aufweisen, bzw. sie weisen Strukturen auf, die sich als Missing Data Probleme interpretieren und mit dafür geeigneten Methoden lösen lassen.

2 Missing Data und Multiple Imputation

Fehlende Daten sind ein häufig auftretendes Phänomen in der empirischen Forschung. Sie werden üblicherweise in die folgenden drei Fälle eingeteilt (vgl. Little und Rubin, 1987, 2002):

- missing completely at random (MCAR); hiermit werden Strukturen bezeichnet, in denen der Ausfallmechanismus unabhängig von den Stichprobenvariablen ist. Dieser Fall liegt beispielsweise vor, wenn die Ausfallwahrscheinlichkeiten für eine Variable, etwa das Einkommen, unabhängig sowohl von den Ausprägungen dieser Variablen selbst als auch von anderen Variablen wie etwa Alter oder sozialem Status sind. Leider ist die Annahme eines MCAR Ausfalles in den meisten Fällen unrealistisch restriktiv.

- missing at random (MAR); dies kennzeichnet Fälle, in denen die Ausfälle durch die beobachteten Daten erklärt werden können. In diesem Fall werden die Antwortausfälle bei einer Variablen, z.B. dem Einkommen, durch andere, vollständig beobachtete Variablen, wie beispielsweise Geschlecht, Alter oder Sozialstatus beeinflusst.

- not missing at random (NMAR); hiermit bezeichnet man Strukturen, in denen die Ausfälle durch die nicht beobachteten Variablen selbst erklärt werden. Dieses Phänomen tritt etwa bei Auskünften zu Einkommen und Vermögenswerten auf, wobei die Antwortneigung in den unteren wie den oberen Einkommensbereichen häufig abnimmt, bei niedrigen/hohen Variablenwerten also mehr fehlende Daten zu verzeichnen sind.

Zur Behandlung von Missing Data Problemen stehen mehrere Vorgehensweisen zur Verfügung (für einen Überblick siehe Rässler und Riphahn, 2006, für Details siehe Little und Rubin, 1987, 2002). Zum einen besteht die Möglichkeit, die Datensätze mit fehlenden Daten ganz oder teilweise zu entfernen, die so genannte *complete/available case analysis*. Dieses Vorgehen reduziert jedoch die verfügbare Information meist erheblich und kann außerdem zu verzerrten Schätzungen führen. Um Fälle von Unit Nonresponse zu korrigieren, werden häufig *Gewichtungen* vorgenommen, indem Gewichte für die vollständig beobachteten Datensätze auf Basis von Hintergrundinformationen über alle Einheiten der Umfrage (Stichprobengewichte) oder der zugrundeliegenden Grundgesamtheit (Populationsgewichte) berechnet werden. *Modellbasierte Verfahren* unterstellen i. allg. parametrische Modelle für die unvollständigen Daten und schlussfolgern auf Basis der Likelihood-Funktionen

bzw. der a posteriori Verteilungen unter diesem Modell. Ein weiteres Vorgehen zur Lösung des Problems fehlender Daten ist die Verwendung von Methoden der *einfachen (single)* bzw. *mehrfachen (multiple) Ergänzung (Imputation)*, die für jeden fehlenden Wert einen oder mehrere Werte ergänzen und so komplettierte Datensätze generieren. Während einfache Imputation hierbei jeweils nur einen Wert ergänzt, beispielsweise den Mittelwert einer beobachteten geeigneten Untergruppe, werden bei mehrfacher Imputation jeweils mehrere Werte für jeden fehlenden Wert ergänzt und so der Unsicherheit bzgl. der fehlenden Daten Rechnung getragen. Derart mehrfach ergänzte Datensätze können wie vollständige Datensätze mit herkömmlichen statistischen Methoden ausgewertet werden. Bei einfach ergänzten Datensätzen führt diese Herangehensweise i. allg. zu unterschätzten Varianzen und zu zu signifikanten p-Werten, dort sind jeweils angepasste Varianzschätzer zu verwenden bzw. ggf. zu entwickeln. Der Charme der Multiplen Imputation (MI) besteht deshalb vor allem darin, dass nach sorgfältig durchgeführter (und oft methodisch durchaus anspruchsvoller) Ergänzung die Datensätze von anderen Personen und zu beliebigen Zeitpunkten mit den üblichen Analyseverfahren statistisch valide ausgewertet werden können, ohne die Tatsache des Datenausfalls bzw. der Ergänzung explizit im Schätzmodell berücksichtigen zu müssen.

Die Verfahren der Multiplen Imputation gehen zurück auf Rubin (1978c, 1987, etc.). Es handelt sich um Monte Carlo Methoden, die für jeden fehlenden Wert m simulierte Werte ergänzen, wobei $m > 1$ gilt, jedoch meist relativ klein sein kann etwa $m = 5$ oder $m = 10$. Die ergänzten Datensätze werden dann mit Standardverfahren ausgewertet und anschließend die Resultate zusammengefasst, um valide Schätzungen zu erhalten, die die Unsicherheit aufgrund der fehlenden Daten widerspiegeln (siehe Abbildung 1).

Die zugrundeliegende Theorie der multiplen Imputation basiert auf der Bayes Statistik, die sich ergebende Inferenz ist im allgemeinem auch vom frequentistischen Standpunkt her gültig. Die Generierung der m Werte erfolgt durch unabhängige Ziehungen aus der a posteriori Prädiktivverteilung $f(y_{mis}|y_{obs})$ der fehlenden Werte gegeben die beobachteten Werte. Da solche Zufallszüge meist Probleme verursachen, und sich diese Dichtefunktion darstellen lässt als

$$f(y_{mis}|y_{obs}) = \int f(y_{mis}, \psi|y_{obs})d\psi = \int f(y_{mis}|y_{obs}, \psi) \cdot f(\psi|y_{obs})d\psi \quad ,$$

wird häufig ein zweistufiger Ziehungsmechanismus verwendet. Hierbei wird zunächst ein Wert des Parameters ψ aus seiner a posteriori Verteilung gegeben die beobachteten Daten $f(\psi|y_{obs})$ gezogen. In einem zweiten

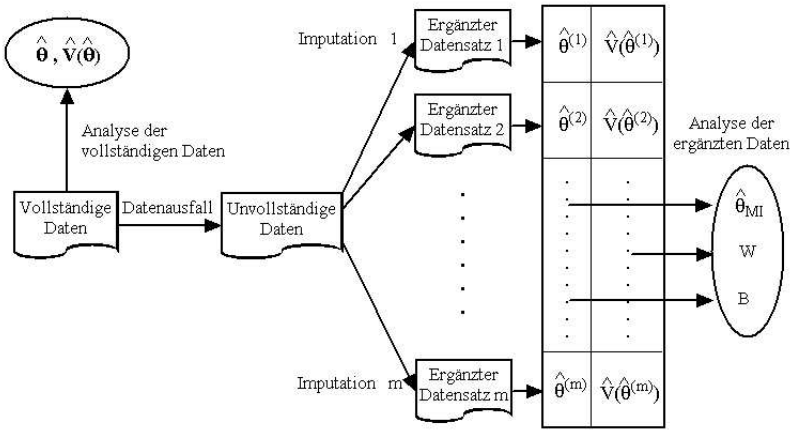

Abbildung 1: Grundprinzip der Multiple Imputation

Schritt wird der fehlende Wert y_{mis} gemäß $f(y_{mis}|y_{obs}, \psi)$ unter Einsatz des Ergebnisses des ersten Schrittes erzeugt.

Da nun wiederum $f(\psi|y_{obs})$ häufig schwierig zu bestimmen ist und nicht gängigen Verteilungsregeln folgt, werden zur Durchführung von mehrfachen Ergänzungen verstärkt *Markov Chain Monte Carlo (MCMC)* Methoden eingesetzt. Ein Spezialfall sind *Data Augmentation* Verfahren (siehe Tanner und Wong, 1987), die eine stochastische Sequenz $\left\{y_{\text{mis}}^{(t)}, \psi^{(t)} : t = 1, 2, \ldots\right\}$ liefern, deren stationäre Verteilung die gesuchte Verteilung $f(y_{\text{mis}}, \psi|y_{\text{obs}})$ ist. Dies wird wiederum durch ein iteratives Verfahren realisiert. Für einen aktuellen Wert $\psi^{(t)}$ des Parameters wird ein Zufallszug $y_{\text{mis}}^{(t+1)}$ der fehlenden Daten, wie bereits beschrieben, aus der bedingten Prädiktivverteilung $f(y_{\text{mis}}|y_{\text{obs}}, \psi^{(t)})$ gezogen. Unter Verwendung dieses neuen Wertes wird nun ebenfalls ein neuer Wert $\psi^{(t+1)}$ für ψ aus seiner a posteriori Verteilung $f(\psi|y_{\text{obs}}, y_{\text{mis}}^{(t+1)})$ der vollständigen Daten gezogen. Züge aus letzterer Verteilung sind meist deutlich einfacher durchzuführen als die vorher beschriebenen Züge aus der a posteriori Verteilung $f(\psi|y_{obs})$ gegeben die beobachteten Daten. Unter der Annahme, das t hinreichend groß ist und die Kette konvergiert, können m unabhängige Züge aus solchen Ketten als multiple Imputationen für Y_{mis} aus seiner a posteriori Prädiktivverteilung $f(y_{\text{mis}}|y_{\text{obs}})$ verwendet werden.

Soll ein unbekannter Parameter θ, der mit ψ übereinstimmen kann aber nicht muss, aus dem durch Multiple Imputation ergänzten Datensatz geschätzt werden, sollten für diesen Schätzer $\widehat{\theta}$ bei vollständigen Daten die Normalverteilung oder die t-Verteilung zumindest approximativ als Referenzverteilung gelten.

Auf Basis der m mehrfach ergänzten Datensätze ergeben sich m Schätzwerte $\widehat{\theta}^{(i)}$ und m Varianzschätzer $\widehat{\mathrm{Var}}(\widehat{\theta}^{(i)})$, $i = 1, \ldots, m$. Der Punktschätzer für θ ergibt sich dann als arithmetisches Mittel $\widehat{\theta}_{MI} = m^{-1} \sum_{i=1}^{m} \widehat{\theta}^{(i)}$, die Varianz als $T = W + (1 + m^{-1})B$ mit

$$B = \frac{1}{m-1} \sum_{i=1}^{m} (\widehat{\theta}^{(i)} - \widehat{\theta}_{MI})^2 \quad \text{als Varianz zwischen den Ergänzungen}$$

und

$$W = \frac{1}{m} \sum_{i=1}^{m} \widehat{\mathrm{Var}}(\widehat{\theta}^{(i)}) \quad \text{als Varianz innerhalb der Ergänzungen.}$$

Der Term $(1 + m^{-1})$ stellt dabei einen Korrekturterm für endliche m dar, der die Varianz im Vergleich zur klassischen Streuungszerlegung $T = B+W$ vergrößert. Die nun folgenden Beispiele sollen zeigen, wie insbesondere die mehrfache Ergänzung für diverse Probleme der Arbeitsmarktforschung eingesetzt werden kann und sich damit teilweise ganz neue Analysemöglichkeiten erschließen.

3 Nutzung von Missing Data Techniken in Projekten des IAB

3.1 Rechtszensierte Lohndaten in der IAB-Beschäftigtenstichprobe

Die Beschäftigtenstichprobe des IAB ist eine zentrale Erhebung zur Erwerbstätigkeit in Deutschland (zu Einzelheiten vgl. Bender *et al.*, 1996). Sie basiert auf einer zweiprozentigen Stichprobe aller sozialversicherungspflichtig Beschäftigten und repräsentiert ungefähr 80% aller Erwerbstätigen in Deutschland. Da es sich um eine Stichprobe aus prozessproduzierten Daten handelt, liegt hier Antwortverweigerung im klassischen Sinn nicht vor. Eines der vordringlichsten mit ihr verbundenen Probleme ist allerdings, dass die Einkommen der Beschäftigten lediglich bis zur Beitragsbemessungsgrenze erfasst werden. Das bedeutet, dass die erfassten Löhne rechtszensiert sind, vgl. Abbildung 2.

Für eine Reihe von Analysen ist eine Information über die Einkommensverteilung im oberen Lohnbereich jedoch unabdingbar. Um dieses Analysepotential zu erschließen, werden die Löhne, die oberhalb der Beitragsbemes-

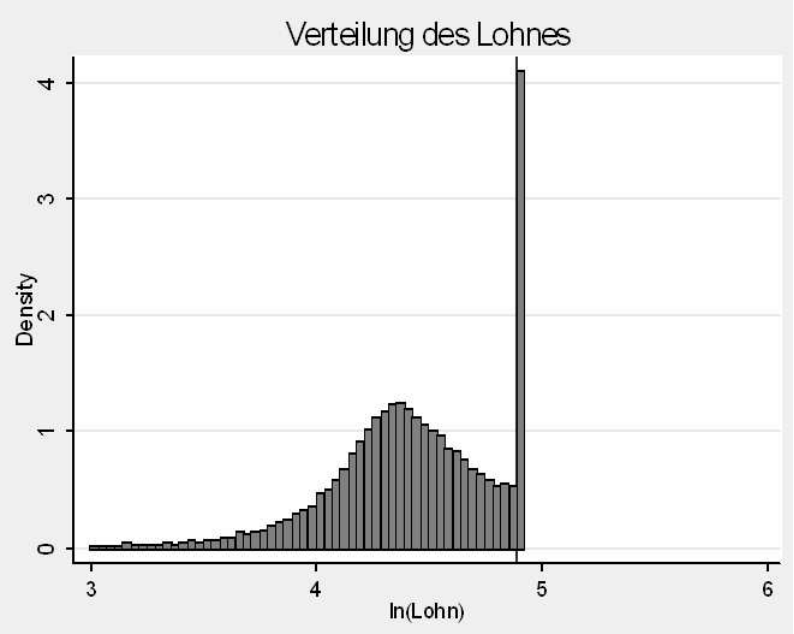

Abbildung 2: Verteilung der Löhne in der IAB-Beschäftigtenstichprobe

sungsgrenze liegen, als Missing Data betrachtet. Sie sind *not missing at random*, da ihr Fehlen auf ihre Ausprägung zurückzuführen ist, und *missing by design*, da sie nicht erhoben werden. Im Rahmen der Forschungsprojekte des IAB zur Analyse von Lohnungleichheiten zwischen Männern und Frauen wurde als erstes ein neues Imputationsverfahren entwickelt und die Werte oberhalb der Beitragsbemessungsgrenze nach einem MCMC Verfahren ergänzt (Gartner und Rässler, 2005).

Betrachtet werden die logarithmierten Tageslöhne von Vollzeitbeschäftigten, die für ein Individuum i für $i = 1, \ldots, n$ durch

$$y_i^* = x_i'\beta + \varepsilon_i \qquad \text{mit} \quad \varepsilon_i \overset{iid}{\sim} N(0; \varsigma^{-2})$$

beschrieben sind. Der Vektor x enthält alle relevanten, den Lohn erklärenden Variablen. Die Löhne $y_{obs} = y_i^*$ werden nur dann beobachtet, wenn der Lohn unter die Beitragsbemessungsgrenze a fällt. Liegen Löhne über a oder sind sie gleich a, wird a anstelle von y_i^* beobachtet:

$$y_i = \begin{cases} y_{obs} & \text{, wenn } y_i^* \leq a \\ a & \text{, wenn } y_i^* > a\,, \end{cases} \tag{1}$$

damit ist die beobachtbare Variable y_i rechtzensiert. Für Löhne oberhalb von a werden nun Schätzungen für die wahren Löhne z imputiert. z ist folglich

eine gestutzte Variable aus $(a;\infty)$ mit der bedingten Verteilung

$$f(z|y,\beta,\varsigma^2) = \frac{f_N(z|x'\beta,\varsigma^{-2})}{1-\Phi(\varsigma a-\varsigma x'\beta)} \quad .$$

Daneben ergeben sich (vgl. Chib, 1992) die a posteriori Verteilung der vollständigen Daten

$$f(\beta|y,z,\varsigma^2) = f_N(\beta|\widehat{\beta}_z,\varsigma^{-2}(X'X)^{-1})$$

und

$$f(\varsigma^2|y,z,\beta) = f_G\left(\varsigma^2\Big|\frac{n}{2}\sum_{i=1}^{n}\frac{(y_{zi}-x_i'\beta_z)^2}{2}\right) \quad ,$$

wobei $y_z = (y_{obs}, z)$, $\widehat{\beta}_z = (X'X)^{-1}X'y_z$, X die Matrix der erklärenden Variablen und N bzw. G die Normal- bzw. Gammaverteilung kennzeichnen. In einem zweistufigen Data Augmentation Verfahren werden dann zunächst fehlende Werte $z_i^{(t)}$ aus der gestutzten Verteilung $N(x_i'\beta^{(t)},\varsigma^{-2(t)})$ gezogen, wobei Startwerte $\beta^{(0)}$ und $\varsigma^{2(0)}$ aus einer Maximum Likelihood Tobit Schätzung verwendet werden. Ein zweiter Schritt dient der Ziehung von $\varsigma^{2(t+1)}$ und $\beta^{(t+1)}$. 11.000 Replikationen und das Speichern der Werte $(z_i^{2000}, z_i^{3000}, \ldots, z_i^{11000})$ führt zu insgesamt zehn vervollständigten Datensätzen. Weitere Einzelheiten der mehrfachen Ergänzung und die Ergebnisse der Lohnungleichheitsanalysen können Gartner und Rässler (2005) entnommen werden. Auch die Analyse der Determinanten ausbildungsinadäquater Beschäftigung mit Hilfe von Einkommensfrontierfunktionen wird durch diese ergänzten Datensätze überhaupt erst möglich. Erstmals gelingt es, die Reichhaltigkeit der Ergebnisse zur Ausbildungsinadäquanz mit dem Einkommensverhältnis-Maß von Jensen (2001) darzustellen (zu den Ergebnissen siehe Jensen *et al.*, 2006).

3.2 Fragebogensplitting bei der Erhebung des gesamtwirtschaftlichen Stellenangebots

Ein weiteres Anwendungsgebiet für Missing Data Techniken in der Arbeitsmarktforschung des IAB findet sich im Bereich der Erhebung des gesamtwirtschaftlichen Stellenangebots. Für Analysen über den Ausgleich von Angebot und Nachfrage auf dem Arbeitsmarkt ist es notwendig, über beide Seiten ausreichende Informationen zur Verfügung zu haben. Während die Angebotsseite (Arbeitslose) jedoch hinreichend detailliert dokumentiert ist, zeigen sich auf

der Seite der Arbeitsnachfrage (offene Stellen) einige Defizite bezüglich Umfang und Struktur der verfügbaren Daten. Da die freiwilligen Meldungen der Arbeitgeber beträchtlich variieren und längst nicht alle offenen Stellen den Arbeitsagenturen gemeldet werden, sind zusätzliche regelmäßige Betriebsbefragungen zur allgemeinen Arbeitskräftenachfrage unabdingbar.

Diese Betriebsbefragungen werden vom IAB seit 1989 bzw. 1992 (West- bzw. Ostdeutschland) jeweils im vierten Quartal eines Jahres durchgeführt. Im Jahr 2004 beteiligten sich hieran ca. 12000 Betriebe, die im Westen 5,6%, im Osten 9,4% aller sozialversicherungspflichtig Beschäftigten repräsentieren. Ermittelt werden auf Basis eines iterativen Hochrechnungsverfahrens die Gesamtzahl der zum Befragungszeitpunkt offenen Stellen - differenziert nach unterschiedlichen Merkmalen - sowie die Stellenbesetzungsvorgänge und wechselnde arbeitsmarktpolitische Fragestellungen, wie beispielsweise die Einschätzung der Auswirkungen einer Arbeitszeitverlängerung.

Ein zentrales Problem dieser Betriebsbefragung stellt die rückläufige Teilnahmebereitschaft der Betriebe dar, denn teilweise war nur jeder fünfte Betrieb bereit, an der Befragung teilzunehmen. Eine systematische Nonresponse-Analyse ergab, dass vor allem die Länge des Fragebogens zusammen mit geringen Personalressourcen in den Betrieben für die Nichtteilnahme verantwortlich scheint. Ziel einer Weiterentwicklung des Fragebogens ist daher die Erhöhung der Rückantwortquote der Betriebe durch Reduktion des Beantwortungsaufwands.

Um dies zu erreichen, gleichzeitig jedoch die Qualität der Befragung zu erhalten, werden Methoden des Fragebogensplittings und der Multiplen Imputation angewendet. In Zusammenarbeit mit der University of Michigan wird derzeit ein geeignetes Splitdesign entwickelt, in dem neben einem gemeinsamen Fragenkomplex mit unverzichtbaren Fragen jedem Teilnehmer nur noch ausgewählte Komponenten zur Beantwortung vorgelegt werden, siehe Abbildung 3.

Die Datenstruktur weist nun Lücken auf, die sich aus dem Aufbau des gesplitteten Fragebogens ergeben, die fehlenden Daten sind also wiederum *missing by design*. Sie werden mit Hilfe von Multiple Imputation ergänzt, so dass zuletzt wieder vollständige Datensätze zur Verfügung stehen. Hierbei muss lediglich beachtet werden, dass Variablenkombinationen in einer kleinen Unterstichprobe gemeinsam beobachtet werden, um Identifikationsprobleme zu vermeiden (vgl. Rässler *et al.*, 2002).

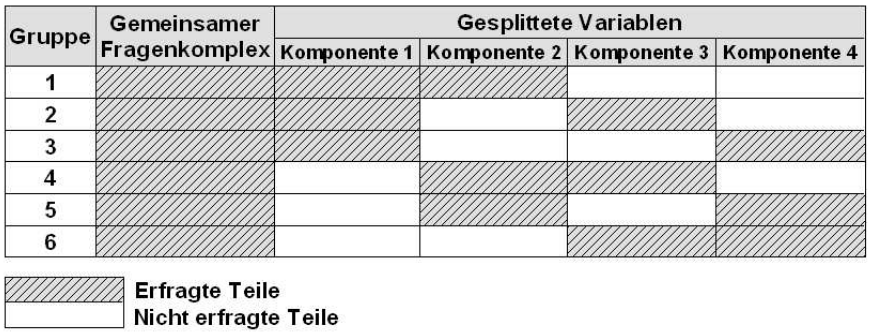

Abbildung 3: Schematische Darstellung des Split-Designs

3.3 Umschätzung von Daten auf Arbeitsagenturebene zu Daten auf Kreisebene und vice versa

Das Problem fehlender Daten ist auch im Zusammenhang mit der Umschätzung von Daten von Agentur- auf Kreisebene virulent. Diese ist notwendig, weil in Deutschland zwei unterschiedliche offizielle regionale Klassifikationen existieren, die beide für die Arbeitsmarktforschung des IAB zugrundegelegt werden. Neben der regionalen Gliederung in Bundesländer, Regierungsbezirke, Landkreise und Gemeinden wird auch die Differenzierung der Bundesagentur für Arbeit in Regionaldirektionen, Arbeitsagenturen und Geschäftsstellen verwendet. Die Grenzen dieser beiden regionalen Klassifikationen schneiden und überlappen sich. Auf der untersten Ebene finden sich die Gemeinden als so genannte *elementare Datenproduktionseinheiten*, siehe Abbildung 4.

Abhängig davon, aus welcher Quelle sie stammen, sind Daten, die für die Arbeitsmarktforschung des IAB verwendet werden, in jeweils einer der beiden Gliederungen auf einer vom Erhebungsschema abhängigen Ebene vorhanden. Beispielsweise liegen Informationen über Investitionssubventionen für Kreise vor, Daten zu Ausgaben und Umfang von Maßnahmen der aktiven Arbeitsmarktpolitik auf Ebene der Agenturen für Arbeit. Wegen der überlappenden Grenzen ist ein einfaches Umrechnen der verfügbaren Informationen auf die jeweils andere Gliederungsstruktur nicht möglich.

Im derzeit noch gängigen Ansatz zur Lösung dieses Problems sucht man eine Variable, die auf der Ebene der elementaren Datenproduktionseinheiten vorhanden ist, etwa die Anzahl der gemeldeten Einwohner. Dann nimmt man stark vereinfachend an, dass die interessierenden Variablen, die nicht auf dieser tiefen Gliederungsebene verfügbar sind, proportional zu der verfügbaren Variablen sind.

Der aktuelle Ansatz des IAB - in Zusammenarbeit mit der Universität

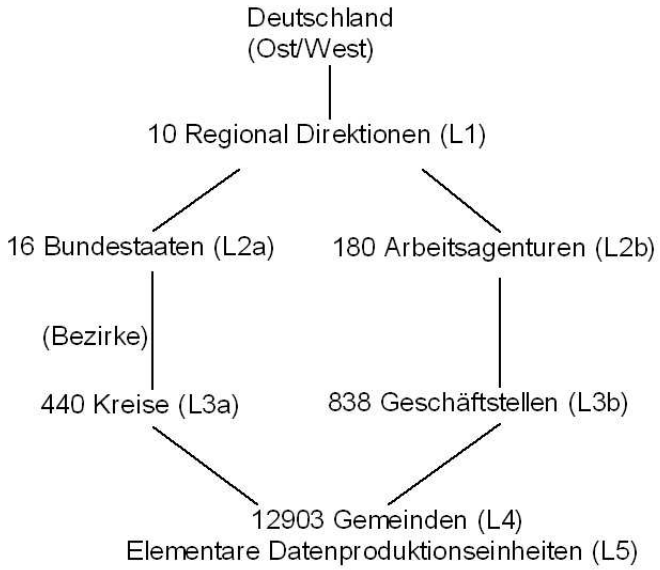

Abbildung 4: Schematische Darstellung der unterschiedlichen regionalen Klassifikationen

Harvard und der Universität von Kalifornien - betrachtet das Problem wiederum als ein Missing Data Problem. Die Daten auf der untersten Ebene können als *missing at random* angenommen und über Multiple Imputation ergänzt werden. Voraussetzung hierbei ist, dass die Randbedingungen, die durch vorhandene Daten auf den höheren Ebenen gegeben sind, eingehalten werden. Das Vorgehen erfolgt wie bei der Ergänzung der rechtszensierten Lohndaten über einen MCMC-Ansatz, der in seinen Grundzügen auf Modelle zur Integration räumlicher Strukturen zurückgeht, die gängigerweise in den Bereichen Astronomie und Astrophysik verwendet werden.

3.4 Umschätzung von Wirtschaftszweigen

Um die einheitliche Erfassung der wirtschaftlichen Tätigkeiten von Unternehmen, Betrieben und anderen statistischen Einheiten in allen amtlichen Statistiken zu gewährleisten, wird in Deutschland die *Klassifikation der Wirtschaftszweige* verwendet. Sie basiert auf der durch EG-Verordnungen verbindliche eingeführten statistischen Systematik der Wirtschaftszweige in der Europäischen Gemeinschaft NACE Rev. 1.1., enthält jedoch eine weitere Gliederungsebene, um die speziell in Deutschland auftretende Strukturen abzubilden. Die aktuelle Ausgabe ist die aus dem Jahr 2003, kurz mit WZ 2003

abgekürzt. Sie ist eine revidierte Version der WZ 93, die ihrerseits die WZ 73 abgelöst hat.

Zwischen 1975 und 2001 verwendete das IAB für seine Arbeit die WZ 73, ab 1999 die WZ 93. Für einen Zeitraum von drei Jahren wurden die beiden Klassifikationen also parallel angewandt. Die Erfassung erfolgt auf der Ebene der fünfstelligen Codezahlen. Der Umstieg von der WZ 73 auf die WZ 93 ist allerdings mit erheblichen Problemen verbunden, da es sich nicht um eindeutige Zuordnungen handelt. Es hat sich gezeigt, dass für die drei Jahre, in denen sowohl das WZ 73 System als auch das WZ 93 System parallel erfasst wurden, ein einzelner WZ 73 Code in 600 verschiedene WZ 93 Codes übergehen kann.

Da bis auf die Daten der Jahre 1999 bis 2001 jeweils nur nach einer der beiden Wirtschaftszweigklassifikationen erhoben wurde, kann auch dieses Problem ein als Missing Data Problem aufgefasst werden, denn jeweils die Informationen nach der zweiten Klassifikation fehlt. Durch Multiple Imputation sollen diese fehlenden Werte im Rahmen einer Doktorarbeit an der Universität Erlangen-Nürnberg ebenfalls geeignet ergänzt werden.

3.5 Produktivitätsschätzungen mit dem IAB-Betriebspanel

Ein klassisches Anwendungsgebiet für Multiple Imputation bietet das IAB-Betriebspanel. Grundlage dieses Panels sind Registerdaten der Bundesagentur für Arbeit, die ungefähr 85% der gesamten Beschäftigung in Deutschland repräsentieren. Aus diesem Register wird eine nach Branchen und Unternehmensgröße geschichtete Stichprobe so gezogen, dass sich größere Unternehmen mit deutlich größerer Wahrscheinlichkeit in der Stichprobe befinden. Die ausgewählten Unternehmen werden in der Regel einmal jährlich durch Interviewer befragt. Die erhobenen Daten enthalten unterschiedliche Informationen zur Beschäftigung in den Betrieben wie Arbeitszeit und Qualifikation der Angestellten, aber auch zum wirtschaftlichen Kontext wie Geschäftsentwicklung und Investitionen.

Auch das IAB-Betriebspanel weist sowohl Unit- als auch Item-Nonresponse auf. Erstere ist wohl vor allem darauf zurückzuführen, dass der Beschaffungsaufwand für die erfragten Informationen dem Betrieb zu groß erscheint bzw. die Informationen nicht zur Verfügung gestellt werden. Item-Nonresponse konzentriert sich im wesentlichen auf die Variablen, die bei Produktivitätsschätzung üblicherweise benötigt werden, um die Größen Produktion, Arbeitskraft und Kapital abzubilden. Hierbei fällt auf, dass der Item-Nonresponse bei kleineren Unternehmen mit geringerem Arbeitskräfteeinsatz und

geringerer technischer Ausstattung wahrscheinlicher ist (vgl. Jensen und Rässler, 2005).

Der häufig eingeschlagenen Weg, nur die vollständig vorhandenen Datensätze in die Analyse einzubeziehen, beinhaltet daher zwei Risiken:

- Zum einen wird durch dieses Vorgehen auf - je nach Analyse - zwischen 20% und 40% der vorhandenen Datensätze verzichtet, die zwar lückenhaft sind, aber ansonsten wertvolle Informationen enthalten.

- Auf der anderen Seite unterstellt diese Herangehensweise, dass die fehlenden Daten MCAR (missing completely at random) sind. Wegen des Zusammenhangs zwischen Unternehmensgröße sowie technischer Ausstattung und Datenausfall, ist jedoch eher davon auszugehen, dass sie (mindestens) MAR (missing at random) sind.

Um trotz der fehlenden Daten eine valide Schätzung von Produktivitäten zu ermöglichen, werden auch hier die Daten mehrfach ergänzt. Dabei tritt der Sonderfall einer quasi doppelten Imputation auf. Neben den fehlenden Angaben weisen die Werte für die Ersatzinvestitionen in auffallend vielen Fällen den Wert Null auf, vermutlich da die Unternehmen den wahren Wert nicht kennen bzw. nicht preisgeben möchten, oder die Beschaffung des Wertes zu mühsam ist. Diese unplausiblen Werte werden ebenfalls als Ausfälle betrachtet und mehrfach ergänzt.

So lange sich die Analysen nur auf maximal zwei Wellen beziehen und die zu ergänzenden Variablen im wesentlichen stetig sind, wird für die Erzeugung der Imputation die kostenlose Software NORM von Schafer (1999) benutzt. Für das vollständige Datenmodell $f(y_{mis}, y_{obs}|\psi)$ werden n unabhängige Beobachtungen aus einer k-dimensionalen Normalverteilung angenommen, d.h. $Y_i \sim N(\mu, \Sigma)$ für $i = 1, \ldots, n$. Ferner ist eine geeignete uninformative a priori Verteilung gegeben durch

$$f(\mu, \sigma) \approx f(\mu) f(\Sigma) \approx c|\Sigma|^{-\frac{k+1}{2}} \propto |\Sigma|^{-\frac{k+1}{2}} \quad ,$$

d.h. μ und Σ werden als a priori unabhängig unterstellt. Mit dieser a priori Verteilung ergibt sich für die a posteriori Verteilung der Parameter mit vollständigen Daten $f(\mu, \Sigma|y)$ das Normal-Wishart-Modell

$$\Sigma|y \sim W^{-1}(n-1, (nS(\bar{y}))^{-1})$$

und

$$\mu|\Sigma, y \sim N(\bar{y}, n^{-1}\Sigma) \quad ,$$

d.h. für Σ eine inverse Wishart Verteilung und für μ eine Normalverteilung. Hierbei bezeichnet S die Stichproben-Kovarianz-Matrix. Die bedingte Prädiktivverteilung $f(y_{mis}|y_{obs}, \mu, \Sigma)$ der fehlenden Daten bei gegebenen beobachtbaren Daten und μ und Σ ist schließlich wieder eine (bedingte) Normalverteilung. Schafer (1997) diskutiert dieses Modell sehr ausführlich.

Wie bei der Ergänzung der rechtszensierten Lohndaten erfolgt der Imputationsprozess als Data Augmentation Verfahren iterativ und in zwei Schritten. In einem ersten Schritt werden für jede Beobachtungseinheit, die fehlende Daten enthält, Werte für diese aus ihrer bedingten Prädiktivverteilung zufällig gezogen. Im zweiten Schritt werden unter Verwendung der vervollständigten Daten Werte für den Erwartungswert und die Varianz-Kovarianz-Matrix berechnet und hierauf Werte der Parameter $\mu^{(t)}$ und $\Sigma^{(t)}$ gemäß ihrer a posteriori Verteilung gezogen. Unter Annahme von Startwerten $\mu^{(0)}$ und $\Sigma^{(0)}$ werden die beiden Schritte wiederholt, bis Unabhängigkeit von den Startwerten hergestellt ist und die Markov-Kette konvergiert. Anschließend werden Werte im Abstand von z. B. 100 gezogenen Werten entnommen, um die fehlenden m Ausprägungen zu erhalten.

Jensen und Rässler (2006) betrachten zu Vergleichszwecken drei unterschiedliche Ansätze:

- Verwendung ausschließlich der beobachteten Daten unter der offensichtlich verletzten MCAR-Annahme;
- Verwendung des vollständigen Datensatzes, wobei fehlende Werte imputiert wurden, die unplausiblen Null-Werte des Kapitaleinsatzes jedoch beibehalten wurden;
- Verwendung des vollständigen Datensatzes mit Imputation der fehlenden Werte und der als Null angegebenen Werte für die Ersatzinvestitionen.

Da es sich um keine Simulationsstudie handelt, kann nicht angegeben werden, welcher der drei Ansätze der Wahrheit am nächsten kommt. Jedoch führen die Ansätze mit imputierten Daten zu plausibleren Ergebnissen als die Analyse nur der vollständigen Datensätze. Kölling und Rässler (2004) stellen beispielsweise eine Reduktion der Produktivitätsdifferenz für das Jahr 2000 zwischen west- und ostdeutschen Firmen von 43% auf 33% bei ergänzten Daten fest. So finden Jensen und Rässler (2006) in einem sehr differenzierten Schätzmodell mit stochastischen Produktionsfrontierfunktionen unerwartet empirische Evidenz für die Wirkung von Flächentarifverträgen in Deutschland. Es zeigen sich mit ergänzten Daten negativen Effekte auf die Produktivität und positive Effekte auf die Effizienz von Firmen.

4 Evaluation aktiver Arbeitsmarktpolitik in der Bundesagentur für Arbeit

Die aktive Arbeitsmarktpolitik der Bundesagentur für Arbeit (BA) umfasst eine große Zahl von Maßnahmen zur Unterstützung der Integration Arbeitsloser oder von Arbeitslosigkeit bedrohter Personen in den Arbeitsmarkt. Hierzu zählen unterschiedlich angelegte Instrumente, die beispielsweise die Qualifikation der Arbeitnehmer verbessern sollen, Arbeitgeber durch Lohnkostenzuschüsse entlasten oder die Aufnahme einer selbständigen Tätigkeit fördern. Gemessen an den Ausgaben entfiel 2004 das größte Gewicht auf die Förderung der beruflichen Weiterbildung (früher Fortbildungs- und Umschulungsmaßnahmen), Arbeitsbeschaffungsmaßnahmen, Eingliederungszuschüsse, sowie Überbrückungsgeld und Existenzgründungszuschüsse.

Für einen effektiven Einsatz dieser Maßnahmen ist eine möglichst präzise Messung ihrer Wirkung unabdingbar. In Zusammenarbeit mit der Harvard Universität entwickeln BA und IAB daher im Rahmen des Projektes TrEffeR (Treatment Effect and PRediction) eine umfassende Wirkungsanalyse der bestehenden arbeitsmarktpolitischen Maßnahmen. Ziel ist zum einen die instrumentübergreifende Evaluation der eingesetzten Maßnahmen, zum anderen die Bereitstellung eines Targeting-Systems, das den Vermittlern in den Agenturen Hilfestellung bei der bestmöglichen Zuweisung von Maßnahmen zu neuen Kunden der BA geben soll.

4.1 Das mikroökonomische Evaluationsproblem

Bei der Evaluation der Maßnahmen der aktiven Arbeitsmarktpolitik gilt es zu ermitteln, ob und wenn ja inwieweit sich die Arbeitsmarktchancen der Maßnahmeteilnehmer durch die Teilnahme an einer bestimmten arbeitsmarktpolitischen Maßnahme verbessert haben. Hierzu wird der Arbeitsmarkterfolg nach der Maßnahme, gemessen an einer geeignet zu spezifizierenden Zielvariablen, mit dem hypothetischen Fall des Erfolgs bei Nicht-Teilnahme (*kontrafaktischer Zustand*) verglichen. Der zugrunde liegende statistische Ansatz wird seit Holland (1986) auch als *Rubins Kausalmodell* bezeichnet (vgl. z.B. Rubin 1974, 1977, 1978a, 1978b).

Im Folgenden bezeichne T allgemein eine arbeitspolitische Maßnahme, der sich Kunde i ($i = 1, \ldots, n$) unterzieht. $P(T_i = 1)$ kennzeichnet die Wahrscheinlichkeit, dass Kunde i an einer Maßnahme teilnimmt. Die Bewertung der Maßnahmen erfolgt anhand geeigneter Kriterien wie Dauer der faktischen Arbeitslosigkeit (Dauer Arbeitslosigkeit einschließlich der Verweildauer in der Maßnahme), Beschäftigungschancen oder Einkommen. Diese Erfolgs-

variablen werden allgemein mit Y bezeichnet. Es gilt

$$T_i = \begin{cases} 1 & \text{falls Person } i \text{ Maßnahmeteilnehmer ist} \\ 0 & \text{sonst} \end{cases} \quad .$$

Somit bezeichnet $Y_i(1)$ den Arbeitsmarkterfolg von Kunde i bei Teilnahme an einer Maßnahme, $Y_i(0)$ den Arbeitsmarkterfolg bei Nicht-Teilnahme. $Y_i(1)$ und $Y_i(0)$ sind hierbei *potentielle Ergebnisse*. Der individuelle kausale Teilnahmeeffekt ergibt sich damit als

$$\tau_i = Y_i(1) - Y_i(0), \quad i = 1...n \quad .$$

Da ein Individuum i entweder an einer Maßnahme teilnimmt oder nicht, ist für die Erfolgsvariable Y an einer Person entweder die Ausprägung $Y(1)|T = 1$ oder die Ausprägung $Y(0)|T = 0$ beobachtbar, niemals jedoch können $Y_i(1)$ und $Y_i(0)$ an ein und derselben Person gemeinsam beobachtet werden. Die unbeobachtbaren, nicht realisierten Ergebnisse werden als kontrakfaktische Zustände bezeichnet und können als Missing Data interpretiert werden. Aufgrund dieses fundamentalen Problems der kausalen Inferenz geht man dazu über, durchschnittliche Maßnahmeeffekte zu betrachten. Der Average Treatment Effect (ATE) ist dabei definiert als der Erwartungswert:

$$\text{ATE} = \text{E}(Y(1)) - \text{E}(Y(0)) \quad .$$

Häufig werden Maßnahmen für ganz bestimmte Zielgruppen entworfen, so dass weniger der allgemeine Maßnahmeeffekt interessiert als der Effekt für die potentiellen Teilnehmer. Der durchschnittliche Maßnahmeeffekt ATE kann hierbei getrennt für die Maßnahmeteilnehmer und die Nicht-Teilnehmer bestimmt werden. Man spricht in diesem Zusammenhang vom *Average Treatment Effect on the Treated (ATET)* und dem *Average Treatment Effect on the non-Treated (ATENT).*

$$\begin{aligned} \text{ATET} &= E(Y(1)|T = 1) - E(Y(0)|T = 1) \\ \text{ATENT} &= E(Y(1)|T = 0) - E(Y(0)|T = 0) \end{aligned}$$

und damit

$$\text{ATE} = \text{ATET} \cdot P(T = 1) + \text{ATENT} \cdot P(T = 0) \quad .$$

Hierbei kennzeichnen $E(Y(0)|T = 1)$ und $E(Y(1)|T = 0)$ die nicht beobachteten Erfolgsgrößen, beobachtbar sind lediglich die Größen $E(Y(0)|T = 0)$

und $E(Y(1)|T=1)$. Ermittelt werden kann somit lediglich:

$$E(Y(1)|T=1) - E(Y(0)|T=0) =$$
$$\underbrace{E(Y(1)|T=1) - E(Y(0)|T=1)}_{\text{ATET}} + \underbrace{E(Y(0)|T=1) - E(Y(0)|T=0)}_{\text{Selektionsverzerrung}} \quad .$$

Erfolgt die Zuweisung von Maßnahmen zufällig, also im Rahmen einer experimentellen Studie, dann sind die potentiellen Ergebnisse von der Zuweisung zur Maßnahme unabhängig, d.h. es gilt $Y(0), Y(1) \perp T$. Aufgrund dieser Unabhängigkeit gilt $E(Y(0)|T=1) = E(Y(0)|T=0) = E(Y(0))$ sowie $E(Y(1)|T=1) = E(Y(1)|T=0) = E(Y(1))$ und somit lassen sich die Ergebnisse der Teilnehmer an einer Maßnahme und die der Nicht-Teilnehmer direkt miteinander vergleichen.

Da Evaluationen jedoch in der Regel auf Basis von Beobachtungsstudien erfolgen, müssen neben der Erfolgsvariablen Y weitere Größen berücksichtigt werden. Der Auswahl eines Maßnahmeteilnehmers aus der Menge aller Arbeitslosen liegt im Regelfall kein Zufallsprozess zugrunde. Vielmehr kann davon ausgegangen werden, dass die Teilnahme an einer Maßnahme von bestimmten Variablen X abhängt, die den Selektionsprozess erklären helfen und auch den Erfolg der Maßnahme beeinflussen können (Rubin, 1977, 1980, Rosenbaum und Rubin, 1983). Diese systematische Unterscheidung von Teilnehmern und Nicht-Teilnehmern an einer Maßnahme wird als *Selektionseffekt* oder *Selektionsverzerrung* bezeichnet. Diese Effekte können zum einen durch den Arbeitslosen selbst (etwa durch höhere Motivation), zum anderen durch den Vermittler (z.B. Auswahl von Teilnehmern nach bestimmten Kriterien) induziert sein.

Um diese Selektionsverzerrungen zu kontrollieren und die kausalen Effekte der Maßnahmen auf den Arbeitsmarkterfolg zu isolieren, werden die Teilnehmer und Nichtteilnehmer an einer Maßnahme T bezüglich dieser weiteren, den Erfolg der Maßnahme möglicherweise beeinflussender Variablen X (beispielsweise Geschlecht, Alter, Qualifikation, Bedingungen auf dem regionalen Arbeitsmarkt,...) bestmöglich angeglichen, die eigentlich wünschenswerte Experimentsituation also nachzustellen versucht. Sofern alle relevanten Variablen berücksichtigt werden, gilt dann $Y(0), Y(1) \perp T|X$. Das Auslassen wichtiger den Förderzugang und den -erfolg erklärender Variabler führt i.allg. zu verzerrten Schätzungen des Maßnahmeeffekts. Ziel ist es daher, in X alle relevanten Pre-Treatment-Faktoren (d.h. die Situation beschreibende Variablen vor und bis zum Zeitpunkt der Maßnahmezuweisung) zu erfassen, Post-Treatment-Faktoren (d.h. die Situation beschreibende Variablen nach dem Zeitpunkt der Maßnahmezuweisung) jedoch auszuklammern, da sie unter Umständen durch die Maßnahme beeinflusst sind.

Der Berechnung der durchschnittlichen Maßnahmeeffekte werden zwei zentrale Annahmen zugrundegelegt:

- Die Annahme der bedingten Unabhängigkeit (Rosenbaum und Rubin, 1983, sprechen von *ignorable treatment assignment*), nach der bei der Bildung der Vergleichsgruppe alle relevanten Merkmale verwendet werden, die Förderzugang und -erfolg determinieren. Die potentiellen Ergebnisse sind dann bei gegebenem X unabhängig von der Zuweisung zu einer Maßnahme und es gilt die oben angesprochene Beziehung

  $$Y(0), Y(1) \perp T|X \quad .$$

  Wird nicht der durchschnittliche Maßnahmeffekt für alle Individuen ATE, sondern der durchschnittliche Effekt für die Teilnehmer ATET berechnet, so ist eine abgeschwächte Form der Bedingung ausreichend. In diesem Fall genügt die Unabhängigkeit der potentiellen Ergebnisse bei Nicht-Teilnahme von der Zuweisung zu einer Maßnahme bei gegebenem X, also $Y(0) \perp T|X$.

- Die SUTVA Annahme (für *stable unit treatment value assumption*), d.h. die Förderung einer Person(engruppe) beeinflusst nicht die Erfolgschancen anderer Personen (keine Interaktionen sowie Substitutions- oder Verdrängungseffekte).

Weiterhin muss zur Bestimmung der Effekte der so genannte *common support* bzw. *overlap* gegeben sein, also gelten:

$$0 < P(T = 1|X) < 1 \quad .$$

Diese Annahme stellt sicher, dass für jede betrachtete Merkmalskombination sowohl Teilnehmer als auch Mitglieder der Kontrollgruppe vorhanden sind (Rosenbaum und Rubin, 1983, sprechen dann insgesamt von *strongly ignorable treatment assignment*). Werden Effekte für die Teilnehmer ermittelt, ist auch für diese Bedingung eine abgeschwächte Form ausreichend, die verlangt, dass bei gegebenem X für jeden Maßnahmeteilnehmer mindestens ein Kontrollelement vorhanden ist, also $P(T = 1|X) < 1$ (vgl. bspw. Imbens, 2004).

Zur Schätzung von Treatment Effekten und Kontrolle von Selektionsverzerrungen steht eine Reihe von Verfahren zur Verfügung. Ein umfangreicher Überblick findet sich unter anderem bei Caliendo und Hujer (2005). Im Rahmen von TrEffeR kommen Matching-Verfahren zum Einsatz, mit deren Hilfe für jeden Teilnehmer ein oder mehrere Nichtteilnehmer gesucht werden, die

bezüglich der Variablen X eine möglichst große Ähnlichkeit aufweisen und als Kontrollgruppe für den Erfolg bei Nichtteilnahme an der Maßnahme (Treatment) fungieren. Damit lässt sich der durchschnittliche Maßnahmeerfolg für Personen mit diesen Charakteristika ermitteln. Allerdings lassen sich diese Befunde nicht auf beliebige andere Personengruppen übertragen.

Die multidimensionale Anpassung von Maßnahmeteilnehmern und Kontrollgruppe führt jedoch bei einer großen Anzahl von erklärenden Variablen zu rechnerischen Problemen. Gut verbreitet und rechnerisch besser umsetzbar sind die so genannten Propensity Score Matching Verfahren (vgl. Rosenbaum und Rubin, 1983). Hierbei werden in Abhängigkeit von den beobachteten Kombinationen an Merkmalen die Förderwahrscheinlichkeiten (Propensity Scores) $P(T = 1|X)$ geschätzt. Rosenbaum und Rubin (1983) zeigen, dass bei Unabhängigkeit der Zielvariablen Y von der Zuweisung T gegeben die Merkmale X dies auch für die Propensity Scores gilt:

$$Y(0), Y(1) \perp T|P(X) \quad .$$

4.2 Die Datenlage

Im Projekt TrEffeR wird zunächst eine Datenbank aufgebaut, die aus den verschiedenen Fachverfahren der BA gewonnen wird und Längsschnittinformationen über Arbeitslosigkeitsepisoden bereitstellt. Die verwendete Datenbasis ist sehr umfangreich, da die Maßnahmeteilnehmer und Nichtteilnehmer nicht nur stichprobenartig, sondern zu 100% erfasst werden. Der für die Analysen berücksichtigte Zeithorizont erstreckt sich zum jetzigen Zeitpunkt über insgesamt fünf Jahre, vom 1. Januar 2000 bis zum 31. Dezember 2004. Die den Selektionsprozess definierenden, im Matching verwendeten erklärenden Variablen X werden für einen Zeitraum von 18 Monaten berücksichtigt. Abzüglich dieser 18-monatigen Historien ergibt sich also der 1. Juli 2001 als faktischer Beginn des Untersuchungszeitraums für die Maßnahmenteilnahme, siehe Abbildung 5.

Der umfangreiche Datensatz stellt eine hinreichend große (und zur Wahrung des Datenschutzes anonymisierte) Anzahl von Maßnahmeteilnehmern und Kontrollgruppenelementen zur Verfügung und ermöglicht so eine detaillierte Betrachtung der Maßnahmeeffekte. Beispielsweise ist es möglich, die Effekte für männliche und weibliche Teilnehmer getrennt auszuwerten und die zeitliche Entwicklung der Maßnahmevergabe und -erfolge zu beobachten, da die Auswertung in Halbjahres-, maximal Jahresschritten erfolgt. Um regionale Unterschiede angemessen zur berücksichtigen, werden die Untersuchungen im Rahmen von TrEffeR auf der Ebene der Arbeitsagenturen durchgeführt. Die verschiedenen Maßnahmen der aktiven Arbeitsmarktpo-

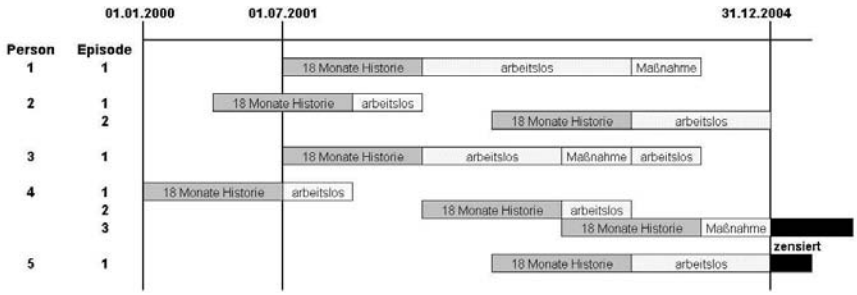

Abbildung 5: Episodenstruktur der Daten

litik wie beispielsweise Eingliederungszuschüsse (EGZ), Förderung der beruflichen Weiterbildung (FbW) oder Beschäftigung schaffende Maßnahmen werden selbstverständlich jeweils gesondert evaluiert, jedoch bei selten vergebenen Maßnahmen geeignet zusammengefasst.

Das auf diesen Ebenen vorgenommene Matching berücksichtigt unter anderem für die Bildung der Kontrollgruppen die folgenden drei Gruppen von erklärenden Variablen:

- Sozio-ökonomische Merkmale wie Alter, Geschlecht, Familienstand, Nationalität, Ausbildung, Beruf, gesundheitliche Einschränkungen.
- Erwerbsbiografie (beschränkt auf 18 Monate) vor Eintritt in die untersuchte Arbeitslosigkeitsepisode mit Merkmalen wie frühere Arbeitslosigkeit, frühere Maßnahmenteilnahme, Sperrzeiten, Krankheiten.
- Zeitstruktur der untersuchten Arbeitslosigkeitsepisode mit Informationen über den Eintrittsmonat in die Arbeitslosigkeit, die Arbeitslosigkeitsdauer bis zum Maßnahmeneintritt, den Kalendermonat des Maßnahmeneintritts.

Um schließlich die berufsbedingten Arbeitsmarktchancen zumindest approximativ abzubilden, werden jeder Episode noch Informationen aus der Statistik der BA über die Anzahl der Arbeitslosen, offene Stellen sowie Einstellungen auf der Basis der dreistelligen Berufskennziffer (BKZ) zugespielt. Welche Variablen sich im Einzelnen dahinter verbergen, kann Tabelle 1 entnommen werden.

Historie	Anzahl der Tage in Arbeitslosigkeit, Anzahl Arbeitslosigkeitsperioden, Anzahl der Tage in Maßnahme, Anzahl Maßnahmen, Anzahl Krankheitstage, Anzahl Krankmeldungen, Anzahl Ortswechsel, Anzahl Tage Leistungsbezug, Anzahl Leistungsbezugsperioden, Anzahl Sperrzeiten
Regionalinformationen	Arbeitslosenquote, Anzahl offene Stellen, Anzahl Einstellungen; jeweils für BKZ 3 Steller
Aktuelle Daten	Alter, Geschlecht, Familienstand, Schulbildung, Berufausbildung, Zugangsgrund, Auswirkung der Behinderung auf die Vermittlung, Behinderungsgrad, Art des Leistungsbezugs Nationalität, Berufsbereich, gewünschte Arbeitszeit, Beginn der Arbeitslosigkeit

Tabelle 1: Matching-Variablen

Daneben besteht an die potentiellen Kontrolleinheiten die Anforderung, dass ihre Arbeitslosigkeitsdauer mindestens so lange sein muss wie die der korrespondierenden Maßnahmeteilnehmer vor Beginn ihrer Maßnahme. Diese Bedingung stellt sicher, dass die ausgewählten Kontrolleinheiten auch zeitlich betrachtet dieselbe Chance hatten, an einer Maßnahme teilzunehmen. Daneben müssen Teilnehmer und Kontrollperson aus der selben Altersklasse stammen, wobei derzeit drei Klassen (unter 25 Jahre, 25 bis 55 Jahre, über 55 Jahre) gebildet wurden. Zuletzt wird zur Eliminierung saisonaler Einflüsse verlangt, dass der Beginn der Arbeitslosigkeitsepisoden eines Teilnehmers von seiner Kontrolleinheit maximal 60 Tage auseinander liegt.

Grundsätzlich kommen als Kontrollpersonen für derartige Effektberechnungen zwei Gruppen in Frage:

- Personen, die nie eine Förderung in Form einer Maßnahme erhalten haben;
- Personen, die nie oder später einer Maßnahme zugewiesen wurden.

In der Regel wird zur Ermittlung von Maßnahmeeffekten die zweite Gruppe als Kontrollgruppe herangezogen. Hierbei ist tendenziell mit positiveren

Maßnahmeeffekten zu rechnen als bei Verwendung der ersten Gruppe, denn diese enthält nur Personen, die nie gefördert wurden, also eher aus eigener Kraft den Weg zurück in den Arbeitsmarkt gefunden haben und somit für günstigere Erfolgsgrößen stehen. Im Rahmen von TrEffeR werden als Kontrollpersonen Individuen herangezogen, die in der betrachteten Arbeitslosigkeitsepisode nicht gefördert wurden, wobei eine Förderung in einer späteren Episode, die durch mehr als sieben Tage von der betrachteten abgegrenzt ist (beispielsweise durch Arbeitsaufnahme) zulässig ist.

4.3 Die Matching-Prozeduren in TrEffeR

Es existieren zahlreiche Spezifikationsmöglichkeiten des Vorgehens beim Matchen von Datensätzen. Welche Herangehensweise optimal ist, hängt von der jeweiligen Datenlage ab. Bei TrEffeR werden im Wesentlichen zwei Ansätze verfolgt:

- ein genetischer Matching-Algorithmus mit Zurücklegen (Programmpaket GenMatch;
  http://sekhon.polisci.berkeley.edu/matching/GenMatch.html);
- ein Nearest Neighbor Propensity Score Matching-Verfahren ohne Zurücklegen mit der zusätzlichen Möglichkeit exakter Matches (Programmpaket MatchIt;
  http://gking.harvard.edu/matchit/).

Die Unterscheidung in mit bzw. ohne Zurücklegen bezieht sich hierbei auf die mehrfache bzw. lediglich einfache Verwendung der Elemente der Kontrollgruppe. Werden die Kontrolleinheiten nur einmal verwendet, ist zu beachten, dass dann die Matching-Reihenfolge bedeutsam wird. Ferner erhöht sich im Vergleich zum Matching mit Zurücklegen der Verzerrung, wohingegen die Varianz geringer ausfallen kann (Dehejia und Wahba, 2002).

Beim gängigen Matching unter Verwendung der Mahalanobis-Distanz gibt die Distanz

$$\mathrm{md}(X_i; X_j) = [(X_i - X_j)' S^{-1} (X_i - X_j)]^{\frac{1}{2}}$$

mit S als Stichproben-Varianz-Kovarianz-Matrix von X an, wie groß die Distanz zweier Beobachtungen in Bezug auf die Werte von X ist. Entsprechend dieser Distanz werden die jeweils Nächsten Nachbarn ermittelt.

Beim genetischen Matching-Algorithmus handelt es sich um ein relativ neues Verfahren, das auf einer Verallgemeinerung der Mahalanobis-Distanz durch eine zusätzliche Gewichtungsmatrix W beruht:

$$\mathrm{d}(X_i; X_j) = [(X_i - X_j)'(S^{-\frac{1}{2}})'WS^{-\frac{1}{2}}(X_i - X_j)]^{\frac{1}{2}} \quad .$$

Hierbei wird die Cholesky-Zerlegung der Stichproben-Varianz-Kovarianz-Matrix S verwendet und W ist eine strikt positive $k \times k$ Diagonalmatrix mit k zu bestimmenden Gewichten. Das Maß gibt an, wie groß die Distanz zweier Beobachtungen i und j in Bezug auf die Werte von X ist, gemäß der die jeweils Nächsten Nachbarn ermittelt werden. Sind alle Gewichte 1, so stimmen md und d überein.

Die Bestimmung der Gewichte erfolgt durch einen evolutionären Algorithmus, der die größte Abweichung zwischen den Gruppen der Maßnahmeteilnehmer und der Nicht-Teilnehmer minimiert. Dabei wird diese größte Abweichung aufgefasst als niedrigster p-Wert aus einer Sequenz durchgeführter Balance-Tests zwischen den beiden Gruppen. Die im Rahmen von TrEffeR maßgeblichen Balance-Tests sind KS-Bootstrap-Tests auf Verteilungsgleichheit. Der Algorithmus variiert in einem iterativen Verfahren die Gewichte der Matrix W stochastisch minimal so, dass die Verteilungen der beiden Gruppen sukzessive angeglichen werden. Ein Abbruch des Verfahrens erfolgt bei TrEffeR, wenn in 10 Generationen keine Verbesserung erzielt wurde, oder wenn die maximale Anzahl von 100 Generationen erreicht ist.

Die Parametrisierung des Verfahrens kann nach verschiedenen Methoden erfolgen. Im Rahmen von TrEffeR wird eine Kombination aus Propensity Score Matching und dem oben beschriebenen Vorgehen gewählt.

Nach einer Erweiterung des Datensatzes um alle quadrierten und logarithmierten Werte der kontinuierlichen Variablen werden zunächst mit Hilfe eines Logit-Modells die Propensity Scores berechnet und der lineare Prädiktor $\widehat{\mu} = X\widehat{\beta}$ extrahiert. Zur Ermittlung der zum Propensity Score orthogonalisierten Variablen wird jede Variable auf $\widehat{\mu}$ und eine Dummyvariable T mit $T = 1$ für die Maßnahmeteilnehmer und $T = 0$ für die Kontrollgruppe regressiert. Die Residuen werden standardisiert und die ersten 15 Hauptkomponenten entnommen. Nach abermaliger Standardisierung des linearen Prädiktors und dieser 15 Hauptkomponenten wird der obige genetische Matching-Algorithmus dann auf die Matrix dieser standardisierten Werte angewendet.

In vergleichenden Betrachtungen mit anderen Matching-Verfahren schneidet diese Vorgehensweise sehr gut ab, da sie die Verteilungen der X Variablen in der Teilnehmergruppe und der Kontrollgruppe üblicherweise am besten angleicht. Sie zeichnet sich jedoch auch durch eine hohe Rechenzeit aus, was angesichts der Menge der zu bewältigenden Daten problematisch ist. Der genetische Matching Algorithmus wird daher als Benchmark verwendet.

Das zweite Verfahren weist deutlich kürzere Rechenzeiten auf. Es handelt sich um ein Nearest Neighbor Propensity Score Matching-Verfahren. Aus-

schlaggebend für die Zuordnung einer Kontrolleinheit zu einem Maßnahmeteilnehmer ist hierbei die kürzeste Distanz zwischen den Propensity Scores, die zuvor mittels eines Logit-Modells geschätzt wurden. Das Matching erfolgt ohne Zurücklegen. Ausführliche Vergleiche beider Verfahren zeigen eine leichte Unterlegenheit von MatchIt gegenüber dem aufwändigeren GenMatch, jedoch häufig relativ ähnliche Ergebnisse. Beide Verfahren laufen unter der Freeware R und werden über SQL-Abfragen von der TrEffer-Datenbank mit "Matchaufträgen" bedient. Solche Matchaufträge bestehen aus den, unter den bereits genannten Restriktionen zusammengestellten Maßnahmeteilnehmern und deren potentiell möglichen Kontrolleinheiten. Die umfangreiche TrEffeR-Datenbank wird also in entsprechende Gruppen aufgeteilt oder besser segmentiert, dies erfolgt nach Agenturen, Geschlecht, Maßnahmenart, Maßnahmeneintritt, sowie teilweise noch nach geplanter Maßnahmedauer. Bei einem Datenbestand von zwischen 35 und 40 Mio Episoden, wie sie Abbildung 5 darstellt, und daraus abgeleiteten ca. 40 Tsd. Matchaufträgen, ist eine weitgehende Automatisierung des gesamten Matchingprozesses unumgänglich.

Zur Sicherstellung des Common Supports werden im Vorfeld alle Maßnahmeteilnehmer von der Evaluation ausgeschlossen, deren Propensity Score größer ist als der größte vorkommende Propensity Score bei den Kontrolleinheiten, da für sie kein entsprechender Matching-Partner gefunden werden könnte.

Da der eigentliche Matching-Prozess ohne Zurücklegen erfolgt, also gegebenenfalls auch einmal weniger geeignete Kombinationen von Teilnehmern und Nicht-Teilnehmern resultieren, werden zusätzlich so genannte *calipers* integriert. Deren Verwendung erhöht wie das Matching mit Zurücklegen die Qualität, indem ungünstige Matches verhindert werden, führt allerdings zu einer höheren Varianz. Bei Verwendung von calipers wird jedem Maßnahmeteilnehmer die Kontrolleinheit zugeordnet, die den nächsten Nachbarn darstellt, jedoch nur, wenn der Abstand der Propensity Scores unterhalb einer vorgegebenen Toleranzschwelle liegt. Bei TrEffeR wird diese Schwelle mit einem Viertel der Standardabweichung der Propensity Scores der Teilnehmer vorgegeben.

4.4 Erste Ergebnisse

Als Ergebnisvariablen werden zur Zeit zwei Größen betrachtet, die die Effektivität der arbeitsmarktpolitischen Maßnahmen abbilden sollen. Hierbei handelt es sich um

- den stichtagsbezogenen *Verbleib in faktischer Arbeitslosigkeit* (VifA)

am Tag t nach Beginn einer Maßnahme und

- die *kumulierte Dauer der faktischen Arbeitslosigkeit* (kDfA) am Tag t nach Beginn einer Maßnahme.

Der Begriff *faktische Arbeitslosigkeit* unterscheidet sich hierbei von dem gesetzlich festgelegten Begriff der Arbeitslosigkeit dadurch, dass die Dauer der Maßnahme einbezogen wird.

Die erste Variable kennzeichnet den Verbleib einer Person in Arbeitslosigkeit zu einem bestimmten Stichtag t, wobei die Stichtage im Abstand von 30 Tagen bis zu maximal 720 Tage nach Maßnahmebeginn angesetzt werden. Die zweite Größe gibt für einen definierten Zeitraum an, wieviele Tage eine Person davon arbeitslos oder in einer arbeitsmarktpolitischen Maßnahme war. Sie kann als relativ strenges Kriterium angesehen werden, da in sie alle Tage nach Maßnahmebeginn einfließen. Die Teilnehmer müssen also die Zeit, die sie in der Maßnahme waren, gegenüber den Kontrollelementen "aufholen".

Auswertungen der Ergebnisse sind im Moment nach folgenden Schnitten - auch in Kombinationen - möglich:

- nach Regionaldirektionen;
- nach Vergleichstypen, also nach Gruppen bezüglich einiger, den Arbeitsmarkt charakterisierender, Merkmale ähnlicher Agenturen (vgl. Blien und Hirschenauer, 2005);
- nach Agenturen;
- nach Maßnahmen;
- für einige Maßnahmen nach deren Dauer;
- nach Geschlecht;
- nach Zugangsjahr und ggf. -halbjahr in Maßnahme.

Die Ergebnisse werden den Arbeitsagenturen in Form von sogenannten "Würfeln" für deren eigene Analysen bereitgestellt und können zur Veranschaulichung graphisch aufbereitet werden. Eine beispielhafte Auswertung dieser Würfel für die Maßnahme Überbrückungsgeld in einer ausgewählten Agentur kann den Tabellen 2 und 3 sowie den Abbildungen 6 und 7 entnommen werden. In den Auswahlblöcken oberhalb der Tabellen 2 und 3 können beliebige Aggregationsstufen eingestellt werden, wobei sich die tiefste Ebene

aus den vorher beschriebenen Schnitten der Matchaufträge ergibt. Abgebildet sind jeweils die Entwicklung der Zielvariablen für die Gruppe der Maßnahmeteilnehmer und die der Nicht-Teilnehmer, sowie der durchschnittliche Maßnahmeeffekt als Differenz der beiden einzelnen Verläufe, inklusive des zugehörigen 95% Konfidenzintervalls. Zu beachten ist, dass in Abweichung von der anfangs dargestellten Definition von Maßnahmeeffekten via

$$\text{ATET} = E(Y(1)|T = 1) - E(Y(0)|T = 1)$$

in dieser Darstellungsform das Vorzeichen umgedreht wird, d.h. die Fördereffekte via VifA (0) - VifA (1) bzw. kDfA (0) - kDfA (1) berechnet werden, um mit einer positiven Förderwirkung auch einen positiven Effekt der betrachteten Maßnahme anzuzeigen.

VglTyp	Alle VglTypen
RD	
Agentur	
Jahr	2002
Halbjahr	1
Geschlecht	männlich
Dauer	ohne Einschr.
MA	ÜG

	Daten				
Tag	Maßnahmeteilnehmer	Kontrollgruppe	Förderwirkung	obere Vertrauensgrenze	untere Vertrauensgrenze
0	1,00	1,00	0,00	0,00	0,00
30	1,00	0,76	-0,24	-0,20	-0,28
60	1,00	0,60	-0,40	-0,35	-0,44
90	1,00	0,51	-0,49	-0,45	-0,53
120	1,00	0,44	-0,56	-0,51	-0,60
150	1,00	0,41	-0,59	-0,54	-0,63
180	0,99	0,40	-0,60	-0,55	-0,64
210	0,07	0,38	0,31	0,36	0,26
240	0,10	0,36	0,26	0,31	0,21
270	0,12	0,36	0,24	0,29	0,19
300	0,14	0,37	0,23	0,29	0,18
330	0,16	0,39	0,23	0,28	0,17
360	0,16	0,42	0,26	0,31	0,20
390	0,15	0,39	0,24	0,29	0,19
420	0,15	0,34	0,18	0,24	0,13
450	0,16	0,31	0,15	0,21	0,10
480	0,15	0,28	0,13	0,18	0,08
510	0,15	0,29	0,13	0,19	0,08
540	0,14	0,29	0,15	0,20	0,10
570	0,17	0,30	0,14	0,19	0,08
600	0,16	0,32	0,16	0,21	0,10
630	0,17	0,32	0,15	0,20	0,10
660	0,18	0,33	0,15	0,20	0,10
690	0,17	0,34	0,17	0,22	0,12
720	0,17	0,34	0,18	0,23	0,12

Tabelle 2: Verbleib in faktischer Arbeitslosigkeit (VifA) Überbrückungsgeld für eine westdeutsche Großstadt

Das Überbrückungsgeld (ÜG) soll für Arbeitslose den Schritt in die Selbständigkeit erleichtern, indem während der Startphase der Gründung für 6 Monate eine Unterstützung in Höhe der Lohnersatzleistung (und zusätzlicher pauschalierter Sozialversicherungsbeiträge) gezahlt wird. Die Abbildungen 6

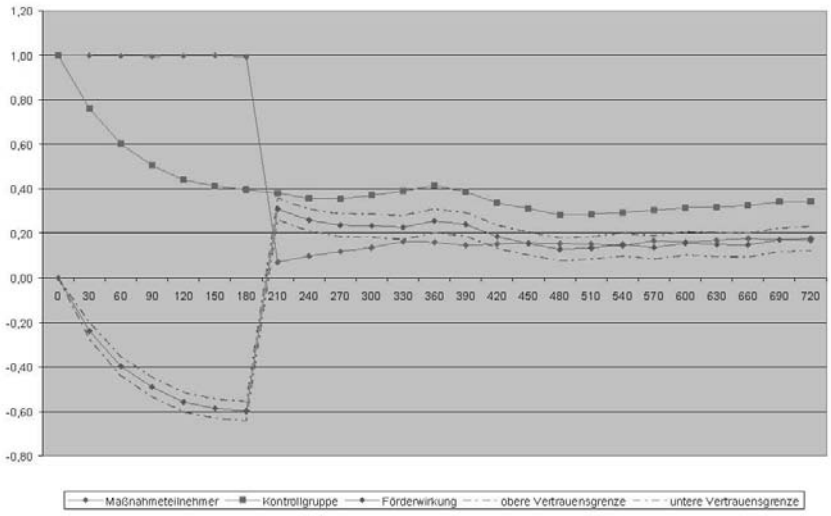

Abbildung 6: Verbleib in faktischer Arbeitslosigkeit (VifA) Überbrückungsgeld für eine westdeutsche Großstadt

VglTyp	Alle VglTypen
RD	
Agentur	
Jahr	2002
Halbjahr	1
Geschlecht	männlich
Dauer	Alle Dauern
MA	ÜG

	Daten				
Tag	Maßnahmeteilnehmer	Kontrollgruppe	Förderwirkung	obere Vertrauensgrenze	untere Vertrauensgrenze
0	0,00	0,00	0,00	0,00	0,00
30	30,00	26,00	-4,00	-3,28	-4,72
60	59,97	46,05	-13,92	-12,19	-15,65
90	89,91	62,56	-27,35	-24,57	-30,13
120	119,84	76,40	-43,44	-39,65	-47,23
150	149,83	89,30	-60,53	-55,77	-65,29
180	179,77	101,32	-78,44	-72,77	-84,11
210	184,14	113,10	-71,04	-64,49	-77,59
240	186,74	124,12	-62,62	-55,16	-70,09
270	189,98	134,96	-55,02	-46,62	-63,42
300	193,90	145,92	-47,98	-38,62	-57,33
330	198,43	157,28	-41,15	-30,82	-51,48
360	203,13	169,61	-33,52	-22,21	-44,84
390	207,70	181,61	-26,08	-13,77	-38,40
420	212,13	192,40	-19,73	-6,38	-33,08
450	216,84	202,16	-14,68	-0,30	-29,06
480	221,55	211,01	-10,54	4,84	-25,92
510	226,21	219,51	-6,70	9,66	-23,06
540	230,65	228,23	-2,42	14,90	-19,74
570	235,24	237,20	1,96	20,28	-16,37
600	240,15	246,58	6,44	25,78	-12,91
630	245,07	256,06	10,99	31,36	-9,38
660	250,32	265,80	15,48	36,90	-5,95
690	255,49	275,82	20,33	42,81	-2,14
720	260,53	286,25	25,72	49,21	2,22

Tabelle 3: Kumulierte Dauer in faktischer Arbeitslosigkeit (kDfA) Überbrückungsgeld für eine westdeutsche Großstadt

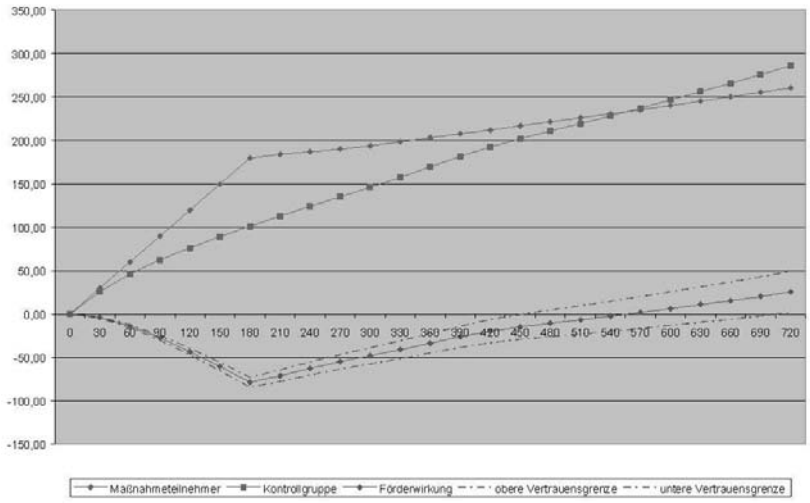

Abbildung 7: Kumulierte Dauer in faktischer Arbeitslosigkeit (kDfA) Überbrückungsgeld für eine westdeutsche Großstadt

und 7 zeigen, dass die Geförderten nach einem anfänglichen "Lock-in"-Effekt deutlich seltener arbeitslos oder in einer Maßnahme als die Vergleichspersonen waren. (Von einem Lock-in-Effekt wird gesprochen, wenn während der Teilnahme an einer arbeitsmarktpolitischen Maßnahme die Suchintensität nach einem neuen Arbeitsplatz reduziert wird.) Nach zwei Jahren (720 Tagen) ist auch der kumulative Effekt der Förderung für diese untersuchte Gruppe signifikant positiv. Die Förderwirkung nimmt im Zeitablauf zwar ab, ist aber auch zwei Jahre nach Beginn der Förderung noch deutlich von Null verschieden.

Eine flächendeckende und systematische Auswertung sämtlicher Ergebnisse steht zum augenblicklichen Zeitpunkt noch aus. Erste Vergleiche, wie etwa Analysen zum Überbrückungsgeld, zeigen jedoch, dass sie konform gehen mit der Studie von Stephan *et al.* (2006), die auf einer zehnprozentigen Stichprobe der TrEffeR-Datenbank basiert und der ein klassisches Propensity Score Matching mit Zurücklegen über die STATA Routine psmatch 2 zugrunde liegt. Stephan *et al.* (2006) nutzen die TrEffeR-Datenbank und zeigen, dass ihre Befunde wiederum mit den Ergebnissen der "Hartz-Evaluation" vereinbar sind, d.h. mit den Ergebnissen im Rahmen der Evaluation der Maßnahmen zur Umsetzung der Vorschläge der Hartz-Kommission.

4.5 Weiteres Vorgehen

Trotz Verfügbarkeit erster Ergebnisse bleiben zahlreiche Fragen und Erweiterungsmöglichkeiten offen.

Zum einen ist vorgesehen, die Menge der Zielvariablen um weitere Größen zu erweitern. Als zusätzliche Variablen wären beispielsweise das Einkommen denkbar oder - vor allem im Hinblick auf potentielle Beitragszahlungen - die Integration in sozialversicherungspflichtige Beschäftigung. Inwieweit jedoch solche Informationen in Verfahren des operativen Geschäftes integriert werden dürfen, bleibt noch mit dem Datenschutz abzuklären.

Daneben ist eine Verlängerung des Beobachtungszeitraums avisiert. Der momentan verfügbare Zeitraum von fünf Jahren stellt insbesondere bei langen Maßnahmen ein Problem dar, da nach Abzug der 18 Monate zum Aufbau der Beschäftigungshistorie und bis zu weiteren 36 Monaten Maßnahmedauer kaum Spielraum für die Beobachtung der Entwicklung nach Ende der Maßnahme besteht. Insbesondere für die Zielvariable kumulierte Dauer faktischer Arbeitslosigkeit ist dies problematisch, da eben die gesamte Maßnahmedauer von den Teilnehmern aufgeholt werden muss, um mit den Nicht-Teilnehmern gleichzuziehen. Kruppe (2006) diskutiert sehr anschaulich die Bedeutung der Wahl eines geeigneten Zielkriteriums insbesondere bei der Evaluation längerfristiger Maßnahmen.

Ferner ermöglicht die momentane Betrachtungsweise lediglich eine Evaluation der Effektivität von Maßnahmen. Um auch hinsichtlich der Effizienz Aussagen treffen zu können, müssen die Kosten einer Maßnahme den Unterstützungsleistungen für Arbeitslose gegenübergestellt werden. Die Implementierung von Kosteninformationen kann demnach ein weiterer für die Zukunft zu berücksichtigender Punkt sein. Gleiches gilt für die Betrachtung makroökonomischer Aspekte wie Substitutions- und Verdrängungseffekte.

Zuletzt ist noch die Umsetzung der Erkenntnisse für das operative Geschäft der BA vorzunehmen. Zum einen ist die Entwicklung eines Prognosesystems zur Unterstützung der Vermittler in den Agenturen ein zentrales Anliegen, um diese bei der Auswahl von geeigneten Maßnahmen für ihre Kunden bestmöglich zu unterstützen. Dass die persönliche Arbeit des Vermittlers vor Ort hierbei durch ein solches Werkzeug nicht ersetzt werden kann, sondern nur eine weitere Entscheidungshilfe auf Basis einer im Normalfall nicht überschaubaren Datenvielfalt bereitgestellt werden soll, ist hierbei selbstredend. Zum anderen besteht die Möglichkeit, auf Basis der Evaluationsergebnisse die bisherige Produktpolitik und -vielfalt auf Effektivität hin zu überprüfen und zu überdenken.

5 Schlussbemerkungen

Missing Data sind ein allgegenwärtiges Datenproblem, sei es bei Stichprobenerhebungen oder auch bei prozessproduzierten Daten. Ihre Behandlung erfordert ein sorgfältiges und auf die jeweilige Situation abgestimmtes Vorgehen und wurde durch die jüngeren Entwicklungen im Bereich der "Computational Statistics" deutlich verbessert, wenn nicht sogar erst ermöglicht.

Exemplarisch wurden einige Probleme aus der Arbeitsmarktforschung des IAB skizziert und Lösungswege über deren Betrachtung als Missing Data Probleme aufgezeigt bzw. vorgestellt.

Auch die wirtschaftspolitisch bedeutsame und in vergangener Zeit verstärkt eingeforderte Evaluation der Instrumente aktiver Arbeitsmarktpolitik lässt sich als Missing Data Problem auffassen, wenn das Konzept der potentiellen Ergebnisse zugrundegelegt wird. Nach den Hartz-Evaluationen unternimmt nun auch die Bundesagentur für Arbeit selbst mit dem Projekt TrEffeR einen ersten großen Schritt, um die aktive Arbeitsmarktpolitik kritisch zu durchleuchten. Die gewählte Konzeption ermöglicht es ihr dabei, diese Evaluation nicht als einmalige Aktion, sondern als kontinuierlichen Prozess zur Optimierung ihrer Arbeit zu begreifen.

Literatur

Bender, S., Hilzendegen, J., Rohwer, G., Rudolph, H. (1996). Die IAB-Beschäftigtenstichprobe 1975-1990. Beiträge zur Arbeitsmarkt- und Berufsforschung 197, Institut für Arbeitsmarkt- und Berufsforschung, Nürnberg.

Blien, U., Hirschenauer, F. (2005). Regionale Arbeitsmärkte: Welche Arbeitsagenturen sind vergleichbar?. IAB-Kurzbericht 18/2005, Institut für Arbeitsmarkt und Berufsforschung.

Caliendo, M., Hujer, R. (2005). The Microeconometric Estimation of Treatment Effects - An Overview. Discussion Paper No. 1653, Forschungsinstitut zur Zukunft der Arbeit, Bonn.

Chib, S. (1992). Bayes Inference in the Tobit Censored Regression Model. *Journal of Econometrics* **51** 79-99.

Dehejia, R.H., Wahba, S. (2002). Propensity Score-Matching Methods for Nonexperimental Causal Studies. *The Review of Economics and Statistics* **84(1)** 151-161.

Gartner, H., Rässler, S. (2005). Analyzing the Changing Gender Wage Gap based on Multiply Imputed Right Censored Wages. IAB Discussion Paper 5/2005.

Hartley, H.O., Hocking, R.R. (1971). The Analysis of Incomplete Data. *Biometrics* **27** 783-808.

Holland, P.W. (1986). Statistical and Causal Inference. *Journal of the American Statistical Association* **81** 945-960.

Imbens, G.W. (2004). Nonparametric Estimation of Average Treatment Effects under Exogeneity: a Review. *The Review of Economics and Statistics* **86(1)** 4-29.

Jensen, U. (2001). *Robuste Frontierfunktionen, methodologische Anmerkungen und Ausbildungsadäquanzmessung.* Peter Lang, Frankfurt.

Jensen, U., Rässler, S. (2005). Where Have all the Data Gone? Stochastic Production Frontiers with Multiply Imputed German Establishment Data. IAB Discussion Paper 15/2005.

Jensen, U., Rässler, S. (2006). The Effects of Collective Bargaining on Firm Performance: New Evidence Based on Stochastic Production Frontiers and Multiply Imputed German Establishment Data. eingereicht als IAB Discussion Paper.

Kalton, G. (1983). Compensating for Missing Survey Data. Research Report Series, Institute for Social Research, University of Michigan, Ann Arbor.

Kruppe, T. (2006). Die Förderung beruflicher Weiterbildung Arbeitsloser im Spiegel von Monitoring und Evaluation. *Zeitschrift für Evaluation* **1/2006** erscheint.

Kölling, A., Rässler (2004). Editing and Multiply Imputing German Establishment Panel Data to Estimate Stochastic Production Frontier Models. *Zeitschrift für Arbeitsmarktforschung* **3/2004** 306-318.

Little, R.J.A. (1976). Inference about Means from Incomplete Multivariate Data. *Biometrika* **63** 593-604.

Little, R.J.A., Rubin, D.B. (1987, 2002). *Statistical Analysis with Missing Data.* Wiley, New York.

Madow, W.G., Olkin, I., Rubin, D.B. (1983). *Incomplete Data in Sample Surveys.* Academic Press, New York.

Rässler, S., Koller, F., Mäenpää, C. (2002). A Split Questionnaire Design applied to German Media and Consumer Surveys. Proceedings of the International Conference on Improving Surveys, ICIS 2002, Copenhagen.

Rässler, S., Riphahn, R. (2006). Survey Item Nonresponse and its Treatment. *Allgemeines Statistisches Archiv* **90** 213-228.

Rosenbaum, P.R., Rubin, D.B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* **70** 41-55.

Rubin, D.B. (1972). A Non-Iterative Algorithm for Least Squares Estimation of Missing Values in any Analysis of Variance Design. *Applied Statistics - Journal of the Royal Statistical Society, Series C* **21** 136-141.

Rubin, D.B. (1974). Characterizing the Estimation of Parameters in Incomplete-Data Problems. *Journal of the American Statistical Association* **69** 467-474.

Rubin, D.B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology* **66** 688-701.

Rubin, D.B. (1976). Inference and Missing Data. *Biometrika* **63** 581-590.

Rubin, D.B. (1977). Assignment to Treatment Group on the Basis of a Covariate. *Journal of Educational Statistics* **2** 1-26, Printer's correction note 3, 384.

Rubin, D.B. (1978a). Bayesian Inference for Causal Effects: The role of Randomization. *The Annals of Statistics* **7** 34-58.

Rubin, D.B. (1978b). Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies. *Journal of the American Statistical Association* **74** 318-328.

Rubin, D.B. (1978c). Multiple Imputation in Sample Surveys - a phenomenological Bayesian approach to nonresponse. In *Proceedings of the Survey Research Methods Sections of the American Statistical Association* 20-40.

Rubin, D.B. (1980). Discussion of Randomization Analysis of Experimental Data in the Fisher Randomization Test by Basu. *Journal of the American Statistical Association* **75** 591-593.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys.* Wiley, New York.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data.* Chapman and Hall, London.

Schafer, J.L. (1999). Multiple Imputation Under a Normal Model. Version 2, Software for Windows 95/98/NT. http://www.stat.psu.edu/ jls/misoftwa.html.

Stephan, G., Rässler, S., Schewe, T. (2006). Wirkungsanalyse in der Bundesagentur für Arbeit: Konzeption, Datenbasis und ausgewählte Befunde. IAB Discussion Paper 07/2006.

Tanner, M.A., Wong, W.H. (1987). The Calculation of Posterior Distributions by Data Augmentation (with Discussion). *Journal of the American Statistical Association* **82** 528-550.

In dieser Reihe sind zuletzt erschienen

Nr.	Autor(en)	Titel	Datum
1/2004	Sabine Hagemann, Werner Sörgel, Eberhard Wiedemann	Vermittlungsgutscheine nach § 421g SGB III - Zwischenergebnisse aus der Begleitforschung zur Vermittlung	9/2004
2/2004	Lutz Bellmann, Vera Dahms, Jürgen Wahse	IAB-Betriebspanel Ost - Ergebnisse der achten Welle 2003 – Teil I: Entwicklung und Struktur der Betriebe und Beschäftigten, Auszubildende	9/2004
3/2004	Lutz Bellmann, Vera Dahms, Jürgen Wahse	IAB-Betriebspanel Ost - Ergebnisse der achten Welle 2003 – Teil II: Personalpolitik, Betriebliche Flexibilität, Weiterbildung	9/2004
4/2004	Lutz Bellmann, Vera Dahms, Jürgen Wahse	IAB-Betriebspanel Ost - Ergebnisse der achten Welle 2003 – Teil III: Wirtschaftliche Lage der Betriebe, Öffentliche Förderung	9/2004
5/2004	Eugen Spitznagel, Susanne Wanger	Mehr Beschäftigung durch längere Arbeitszeiten? Ein Beitrag zu der Diskussion um eine generelle Erhöhung der Arbeitszeit	10/2004
6/2004	IAB-Autorengemeinschaft	Forschung zum SGB II des IAB: Die neuen Forschungsaufgaben im Überblick	12/2004
1/2005	Anja Heinze, Friedhelm Pfeiffer, Alexander Spermann, Henrik Winterhager, Amelie Wuppermann	Vermittlungsgutscheine - Zwischenergebnisse der Begleitforschung 2004 Teil I: Datenstruktur und deskriptive Analysen	3/2005
2/2005	Sabine Dann, Günther Klee, Martin Rosemann	Vermittlungsgutscheine - Zwischenergebnisse der Begleitforschung 2004 Teil II: Typisierung der Arbeitsagenturen	2/2005
3/2005	Anja Heinze, Friedhelm Pfeiffer, Alexander Spermann, Henrik Winterhager	Vermittlungsgutscheine - Zwischenergebnisse der Begleitforschung 2004 Teil III: Mikroökonometrische Wirkungsanalyse	3/2005

4/2005	Reinhard Hujer, Christopher Zeiss	Vermittlungsgutscheine - Zwischenergebnisse der Begleitforschung 2004 Teil IV: Makroökonomische Wirkungsanalyse	2/2005
5/2005	Friedhelm Pfeiffer, Henrik Winterhager	Vermittlungsgutscheine - Zwischenergebnisse der Begleitforschung 2004 Teil V: Kosten-Nutzen-Analyse	2/2005
6/2005	Sabine Hagemann, Werner Sörgel	Vermittlungsgutscheine - Zwischenergebnisse der Begleitforschung 2004 Teil VIa: Implementations- und Strukturanalysen - Private Arbeitsvermittler	7/2005
7/2005	Sabine Hagemann, Werner Sörgel	Vermittlungsgutscheine - Zwischenergebnisse der Begleitforschung 2004 Teil Vb: Implementations- und Strukturanalysen - Tabellenanhang	7/2005
8/2005	Reinhard Hujer, Günther Klee, Alexander Spermann, Werner Sörgel	Vermittlungsgutscheine - Zwischenergebnisse der Begleitforschung 2004 Teil VII: Zusammenfassung der Projektergebnisse	7/2005
9/2005	Regina Konle-Seidl	Lessons learned – Internationale Evaluierungsergebnisse zu Wirkungen aktiver und aktivierender Arbeitsmarktpolitik	2/2005
10/2005	Ch. Brinkmann, J. Passenberger, H. Rudolph, E. Spitznagel, G. Stephan, U. Thomsen, H. Roß	SGB II – Neue Herausforderungen an Statistik und Forschung	2/2005
11/2005	Corinna Kleinert, Hans Dietrich	Aus- und Weiterbildungen im Pflegebereich - Eine Analyse des Eingliederungsprozesses in Erwerbstätigkeit	3/2005
12/2005	Axel Deeke	Kurzarbeit als Instrument betrieblicher Flexiblität - Ergebnisse aus dem IAB-Betriebspanel 2003	3/2005
13/2005	Oliver Falck	Das Scheitern junger Betriebe Ein Überlebensdauermodell auf Basis des IAB-Betriebspanels	3/2005
14/2005	Helmut Rudolph, Kerstin Blos	Schätzung der Auswirkungen des Hartz-IV-Gesetzes auf Arbeitslosenhilfe-Bezieher	4/2005

15/2005	Johann Fuchs, Brigitte Weber	Neuschätzung der Stillen Reserve und des Erwerbspersonenpotenzials für Westdeutschland (inkl. Berlin-West)	5/2005
16/2005	Johann Fuchs, Doris Söhnlein	Vorausschätzung der Erwerbsbevölkerung bis 2050	5/2005
17/2005	Michael Feil, Gerd Zika	Politikberatung mit dem Simulationsmodell PACE-L – Möglichkeiten und Grenzen am Beispiel einer Senkung der Sozialabgaben	5/2005
18/2005	Johann Fuchs, Brigitte Weber	Neuschätzung der Stillen Reserve und des Erwerbspersonenpotenzials für Ostdeutschland (einschl. Berlin-Ost)	6/2005
19/2005	Stefan Schiel, Ralph Cramer, Reiner Gilberg, Doris Hess, Helmut Schröder	Das arbeitsmarktpolitische Programm FAIR - Zwischenergebnisse der Begleitforschung 2004 -	7/2005
20/2005	Lutz Bellmann, Vera Dahms, Jürgen Wahse	IAB-Betriebspanel Ost – Ergebnisse der neunten Welle 2004 – Teil I: Entwicklung und Struktur der Betriebe und Beschäftigten, Auszubildende	7/2005
21/2005	Lutz Bellmann, Vera Dahms, Jürgen Wahse	IAB-Betriebspanel Ost – Ergebnisse der neunten Welle 2004 – Teil II: Personalpolitik, Betriebliche Flexibilität, betriebliche Arbeitszeiten, ältere Arbeitnehmer	7/2005
22/2005	Lutz Bellmann, Vera Dahms, Jürgen Wahse	IAB-Betriebspanel Ost – Ergebnisse der neunten Welle 2004 – Teil III: Innovationen im Betrieb, wirtschaftliche Lage der Betriebe	7/2005
23/2005	Aderonke Osikominu	Eine Analyse der Teilnehmerselektion in die berufliche Weiterbildung auf Basis der Integrierten Erwerbsbiografien (IEB)	9/2005
24/2005	Uwe Blien, Franziska Hirschenauer	Vergleichstypen 2005: Neufassung der Regionaltypisierung für Vergleiche zwischen Agenturbezirke	9/2005
25/2005	Johann Fuchs, Katrin Dörfler	Projektion des Erwerbspersonenpotenzials bis 2050 – Annahmen und Grundlagen	9/2005
26/2005	Axel Deeke	Das ESF-BA-Programm im Kontext der arbeitsmarktpolitischen Neuausrichtung der Bundesagentur für Arbeit – Zur Umsetzung des Programms von 2000 bis Anfang 2005	10/2005

1/2006	Lena Koller, Ulrike Kress, Kerstin Windhövel	Blinde Kuh war gestern – heute ist FIS Das Forschungs-Informations-System – ein neuer Weg wissenschaftlicher Politikberatung	1/2006
2/2006	Susanne Wanger	Erwerbstätigkeit, Arbeitszeit und Arbeitsvolumen nach Geschlecht und Altersgruppe – Ergebnisse der IAB-Arbeitszeitrechnung nach Geschlecht und Alter für die Jahre 1991-2004	1/2006
3/2006	Sarah Heinemann, Hermann Gartner, Eva Jozwiak	Arbeitsförderung für Langzeitarbeitslose - Erste Befunde zu Eingliederungsleistungen des SGB III im Rechtskreis SGB II	2/2006
4/2006	Jan Binder, Barbara Schwengler	Neuer Gebietszuschnitt der Arbeitsmarktregionen im Raum Berlin und Brandenburg – Kritische Überprüfung der bisher gültigen Arbeitsmarktregionen und Vorschläge für einen Neuzuschnitt	2/2006
5/2006	Ch. Brinkmann, M. Caliendo, R. Hujer, St. L. Thomsen	Zielgruppenspezifische Evaluation von Arbeitsbeschaffungsmaßnahmen – Gewinner und Verlierer	2/2006
6/2006	Ch. Gaggermeier	Indikatoren-Modelle zur Kurzfristprognose der Beschäftigung in Deutschland	4/2006
7/2006	St. Schiel, R. Gilberg, H. Schröder	Evaluation des arbeitsmarktpolitischen Programms FAIR - 3. Zwischenbericht	4/2006
8/2006	K. Blos	Die Bedeutung der Ausgaben und Einnahmen der Sozialversicherungssysteme für die Regionen in Deutschland	3/2006
9/2006	A. Haas, Th. Rothe	Regionale Arbeitsmarktströme - Analysemöglichkeiten auf Basis eines Mehrkontenmodells	4/2006
10/2006	J. Wolff, K. Hohmeyer	Förderung von arbeitslosen Personen im Rechtskreis des SGB II durch Arbeitsgelegenheiten: Bislang wenig zielgruppenorientiert	6/2006
11/2006	L. Bellmann, H. Bielenski, F. Bilger, V. Dahms, G. Fischer, M. Frei, J. Wahse	Personalbewegungen und Fachkräfterekrutierung – Ergebnisse des IAB-Betriebspanels 2005	6/2006

12/2006	Th. Rhein, M. Stamm	Niedriglohnbeschäftigung in Deutschland: Deskriptive Befunde zur Entwicklung seit 1980 und Verteilung auf Berufe und Wirtschaftszweige	7/2006
13/2006	B. Rudolph, C. Klement	Arbeitsmarktpartizipation von Frauen im Transformationsprozess - Sozio-ökonomische Realität in den EU-Beitrittsländern Polen, Tschechien und Ungarn	7/2006
14/2006	Th. Rothe	Die Arbeitskräftegesamtrechnung für Ost- und Westdeutschland – Konzeption und ausgewählte Ergebnisse	7/2006
15/2006	R. Konle-Seidl, Kristina Lang	Von der Reduzierung zur Mobilisierung des Arbeitskräftepotenzials	8/2006
16/2006	Johanna Dornette, Marita Jacob	Zielgruppenerreichung und Teilnehmerstruktur des Jugendsofortprogramms JUMP	8/2006
17/2006	Andreas Damelang, Anette Haas	Arbeitsmarkteinstieg nach dualer Berufsausbildung – Migranten und Deutsche im Vergleich	8/2006

Impressum

IAB*Forschungsbericht*
Nr. 18 / 2006

Herausgeber
Institut für Arbeitsmarkt- und Berufsforschung
der Bundesagentur für Arbeit
Weddigenstr. 20-22
D-90478 Nürnberg

Redaktion
Regina Stoll, Jutta Palm-Nowak

Technische Herstellung
Jutta Sebald

Bezugsmöglichkeit
Volltext-Download dieses Forschungsberichtes
unter:
http://doku.iab.de/forschungsbericht/2006/fb1806.pdf

IAB im Internet
http://www.iab.de

Rückfragen zum Inhalt an
Susanne Rässler, Tel. 0911/179-3084,
oder E-Mail: Susanne.Rässler@iab.de