Fehlende Daten bei der Verknüpfung von Prozess- und Befragungsdaten

Ein empirischer Vergleich ausgewählter Missing Data Verfahren

Missing Data due to Record Linkage of Register and Survey Information

An Empirical Comparison of Selected Missing Data Techniques

Gerhard Krug

Zusammenfassung

Die Verknüpfung von Prozess- und Befragungsdaten gewinnt in der empirischen Sozialforschung zunehmend an Bedeutung. Aus Datenschutzgründen können Befragte die Verknüpfung aber ablehnen, weshalb die verbleibende Stichprobe selektiv sein kann. Hier können Missing Data Techniken helfen, eventuelle Selektionsverzerrungen in empirischen Analysen zu korrigieren. Dieses Papier nutzt eine Befragung, in der unter anderem die Zustimmung zur Verknüpfung erbeten wurde, um den Erfolg ausgewählter Missing Data Techniken bei der Ausfallkorrektur im Rahmen einer Fallstudie zu vergleichen. Bei nicht zustimmenden Befragten werden ihre faktisch gegebenen Antworten auf „fehlend" gesetzt, um so pseudo-fehlende Werte auf Basis eines empirischen (im Vergleich zu einem statistisch simulierten) Ausfallmechanismus zu erzeugen. Eine KQ-Regressionsanalyse wird durchgeführt und eventuelle Verzerrungen durch den Datenausfall werden jeweils alternativ durch fallweisen Ausschluss von Beobachtungen mit fehlenden Werten, eine Multiple Imputation

Abstract

Linking register to survey data is becoming more and more important for empirical social science. Due to reasons of data protection the respondents have be asked for their permission to link their data. The resulting sample can therefore be selective. Missing data techniques can be used to correct for any record linkage bias. In this paper I use a survey where participants were asked permission for combining the survey with administrative data (record linkage). Based upon this survey the performance of different missing data techniques is compared. For those who refuse their permission I set their survey answers to missing, creating pseudo-missing data following an empirical relevant but unknown mechanism (rather than a statistical simulation of a missing data process). OLS Regression is performed using casewise deletion, multiple imputation and two versions of Heckman's sample selection model, respectively, to correct for the pseudo-missing data. The results are compared to a regression that is based on the complete data set and that gives us the

(Ergänzung) der fehlenden Werte und durch Selektionskorrektur nach Heckman korrigiert. Die Ergebnisse der Korrekturverfahren werden mit Regressionsanalysen auf Basis der vollständigen Daten verglichen, welche die „wahren" Regressionskoeffizienten liefern. Es zeigt sich in einer Beispielanalyse mit *geringer* Selektivität des Datenausfalles, dass hier alle Korrekturverfahren ähnlich gut abschneiden. In einer zweiten Analyse mit *starker* Selektivität lieferte ausschließlich die Multiple Imputation gute Ergebnisse, jedoch nur, wenn die abhängige Variable keine fehlenden Werte aufwies.

"true" regression parameters. In an empirical example analysis characterized by weak selectivity of the missing data, all missing data techniques performed quite well. In a second example analysis with strong selectivity, it was only multiple imputation that was able to correct for the record linkage bias, given that missing values were present only in one or more independent variables. In the case of strong selectivity and missing values in the dependent variable, none of the missing data techniques eliminated the bias.

1 Einleitung[1]

Standardisierte Befragungen sind ein zentrales Element der empirischen Sozialforschung. Durch die Gründung von Forschungsdatenzentren und die Aufbereitung administrativer Daten zu Scientific Use Files rücken aber auch sogenannte prozessproduzierte Daten in den Blick der Forschung (Wirth/Müller 2004; Allmendinger/Kohlmann 2005). Angesichts der Tatsache, dass die Vorteile beider Datenquellen zum Teil auf unterschiedlichen Gebieten liegen (vgl. Hartmann/Krug 2009), bietet es sich an, die Aussagekraft empirischer Analysen durch ihre Kombination zu erweitern. Eine Möglichkeit besteht in der Datenverknüpfung (Record Linkage), bei der auf der individuellen Ebene Befragungsdaten mit prozessproduzierten Informationen zum selben Individuum angereichert werden.[2] In vielen Fällen gilt allerdings, dass vor einer solchen Verknüpfung die Erlaubnis der betroffenen Personen einzuholen ist. Obwohl erfahrungsgemäß die Zustimmungsbereitschaft der Befragten relativ hoch ist (z. B. Hartmann et al. 2008: 57), wird diese natürlich auch von einem Teil der Befragten verweigert. Diese Personen weisen dann in dem angereicherten Datensatz bei den entsprechenden Variablen fehlende Werte auf.

1 Für wertvolle Hinweise und Verbesserungsvorschläge zu früheren Versionen des Textes danke ich Jörg Drechsler, Hans Kiesl, André Pahnke, Martin Spieß, Gesine Stephan sowie den Herausgebern und zwei anonymen Gutachtern der MDA. Für das Korrekturlesen des Textes danke ich Katrin Drasch und Christiane Spies. Verbliebene Fehler liegen in meiner Verantwortung.

2 Vom Record Linkage ist das statistische Matching (Rässler 2002) zu unterscheiden, bei dem Informationen von aus statistischer Sicht möglichst ähnlichen Individuen miteinander verknüpft werden.

Für die empirische Forschung mit solchen Daten stellt sich damit die Frage des Umgangs mit den fehlenden Werten. Die einfachste Lösung besteht darin, für Analysen alle Fälle mit fehlenden Werten auszuschließen (fallweiser Ausschluss). Dies setzt jedoch einen zufälligen Ausfall der nicht verwendeten Beobachtungen voraus. Ist dies nicht der Fall, werden Schätzungen etwa von Mittelwerten oder Regressionskoeffizienten verzerrt sein. Statistische Verfahren, wie die Multiple Imputation (Ergänzung) fehlender Werte oder eine Selektionskorrektur nach Heckman, versprechen hier Abhilfe. Dabei gehen sie von bestimmten Annahmen über den Datenausfallprozess aus, so dass bei Nichterfüllung dieser Annahmen das Ziel unverzerrter Schätzungen aber eventuell verfehlt wird.

Da diese Annahmen im konkreten Anwendungsfall meist nicht testbar sind, ist die Entscheidung für das eine oder andere Verfahren zum Umgang mit fehlenden Werten (sogenannte Missing Data Verfahren) oft schwierig. Die vorliegende Arbeit prüft im Rahmen einer Fallstudie einige ausgewählte Verfahren und zeigt auf, inwiefern sie bei der Korrektur von Stichprobenausfällen zu unterschiedlichen Ergebnissen führen. Aufgrund des gewählten Analysedesigns kann an ausgewählten Beispielen nicht nur nachvollzogen werden, ob die Verfahren in der Forschungspraxis zum selben, sondern auch zum *richtigen* Ergebnis gelangen. Es wird hierzu eine Befragung genutzt, bei der die Zustimmung zur Verknüpfung mit administrativen Daten abgefragt wurde. Solche empirischen Vergleiche sind etwa geeignet, um die Robustheit empirischer Forschungsergebnisse im Hinblick auf das gewählte Verfahren unter realistischen Anwendungsbedingungen zu analysieren (vgl. z. B. Ridder 1992). Im Unterschied zu vollständig simulierten Daten ermöglicht der echte Datensatz in Kombination mit dem gewählten Analysedesign, das Fehlen von Prozessdaten aufgrund tatsächlicher empirischer Teilnahmeentscheidungen von Befragten nachzuahmen. So liefern die folgenden Analysen im Gegensatz zu Simulationen einen Hinweis, ob die Missing Data Verfahren geeignet sind, den *empirisch* aufgrund bestehender Datenschutzregelungen auftretenden Datenausfall bei der Verknüpfung von Prozess- und Befragungsdaten auszugleichen.

Die vorliegende Arbeit ist hierzu wie folgt aufgebaut. Zunächst werden in Abschnitt 2 Annahmen zu verschiedenen Ausfallmechanismen und mit diesen korrespondierende Verfahren zum Umgang mit fehlenden Werten vorgestellt. Im Anschluss (Abschnitt 3) wird der empirische Vergleich durchgeführt, wobei zunächst die Datenbasis vorgestellt sowie die empirischen Determinanten des Datenausfalles bestimmt werden (3.1). Danach wird das Design des Vergleichs vorgestellt (3.2), die Implementation der Missing Data Verfahren besprochen (3.3) und schließlich werden die Ergebnisse präsentiert (3.4). In Abschnitt 4 erfolgt eine Zusammenfassung und Diskussion der Ergebnisse und Abschnitt 5 schließt mit Schlussfolgerungen aus dem empirischen Vergleich der Verfahren.

2 Ausfallmechanismen und Missing Data Verfahren

Die Zusammenspielung von Prozess- und Befragungsdaten kann unterschiedlichen Zwecken dienen. Im Folgenden wird davon ausgegangen, dass mit den verknüpften Prozess- und Befragungsdaten eine Regressionsanalyse durchgeführt werden soll, wobei mit Y_i die abhängige Variable und mit $\mathbf{X}_i$ der Vektor der j = 1,2,..., J unabhängigen Variablen bezeichnet wird. Von Interesse seien die Koeffizienten der Regression $Y_i = \mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i$ (zur übersichtlicheren Darstellung wird im Folgenden auf den Personenindex i verzichtet), welche im Folgenden als *Analysegleichung* oder *Analysemodell* bezeichnet wird.

Grundsätzlich lassen sich hinsichtlich des Datenausfalls im Allgemeinen und so auch im Fall der Datenverknüpfung drei unterschiedliche Situationen unterscheiden (Rubin 1987; Little/Rubin 1987; Collins/Schafer/Kam 2001): missing at random (MAR), missing completely at random (MCAR) und missing not at random (MNAR). Diese Situationen unterscheiden sich danach, welche Annahmen über die Beziehung zwischen den in der konkreten inhaltlichen Analyse relevanten Variablen und den Determinanten des Ausfallprozesses gerechtfertigt sind (vgl. zum Folgenden Horton/Lipsitz 2001; Horton/Kleinman 2007): Für jede befragte Person kann $\mathbf{X}$ in zwei Komponenten unterteilt werden: $\mathbf{X}^{obs}$ bezeichnet die Variablen ohne fehlende Werte (die beobachte Komponente) und $\mathbf{X}^{mis}$ die Variablen mit fehlenden Werten (die fehlende Komponente), entsprechendes gilt für Y^{obs} und Y^{mis}. Sei $\mathbf{R}$ ein Vektor von j = 1,2,..., J Indikatorvariablen, die für jede x-Variable angeben, ob der entsprechende Wert fehlt (R_j = 1, falls das j-te Element von $\mathbf{X}$ fehlt, R_j = 0 sonst) und sei ϕ der Parametervektor, der den Ausfallprozess (ein Element wird beobachtet oder nicht) kennzeichnet. $P(\mathbf{R} \mid Y, \mathbf{X})$ sei die Wahrscheinlichkeit für das Auftreten eines bestimmten Ausfallmusters.

Nach Little und Rubin (1987) ist missing completely at random (MCAR) definiert als

$$P(\mathbf{R} \mid Y, \mathbf{X}) = P(\mathbf{R} \mid Y^{obs}, Y^{mis}, \mathbf{X}^{obs}, \mathbf{X}^{mis}) = P(\mathbf{R} \mid \phi) \qquad (1)$$

wobei ϕ und β (zu schätzende Parameter) als distinkt angenommen werden. Weniger technisch ausgedrückt bedeutet MCAR: „the process generating missing values bears no statistical relationship (e.g. correlations) with our variables of interest" (Collins/Schafer/Kam 2001: 333). Diese Annahme erscheint jedoch gerade dann, wenn der Ausfall auf Entscheidungen der befragten Individuen beruht, als problematisch. Die missing at random (MAR) Annahme ist dagegen weniger restriktiv und lautet:

$$P(\mathbf{R} \mid Y, \mathbf{X}) = P(\mathbf{R} \mid Y^{obs}, \mathbf{X}^{obs}, \phi) \quad (2)$$

Die MAR-Annahme besagt damit, dass der Ausfallprozess zwar mit den Variablen der interessierenden Analyse zusammenhängt, diese Beziehung aber vollständig von den beobachteten Daten erfasst wird. Graham (2009: 553) spricht daher auch von „conditionally missing at random". Im Fall MCAR wie auch MAR wird der Ausfallprozess (missing data mechanism) als ignorierbar (ignorable) bezeichnet. Dagegen besagt die MNAR-Annahme, dass der Ausfallprozess $P(\mathbf{R} \mid Y, \mathbf{X})$ nicht weiter vereinfacht werden kann, da er auch auf unbeobachteten Daten beruht; er ist nicht ignorierbar (nonignorable) (Little/Rubin 1987):

$$P(\mathbf{R} \mid Y, \mathbf{X}) \neq P(\mathbf{R} \mid Y^{obs}, \mathbf{X}^{obs}, \phi) \quad (3)$$

Es gibt eine Reihe von Verfahren, mit fehlenden Daten umzugehen (Missing Data Verfahren). Diese lassen sich danach unterscheiden, welche der drei genannten Annahmen zum Datenausfall mindestens erfüllt sein muss, damit die Verfahren anwendbar sind.

2.1 Fallweiser Ausschluss: Missing completely at random (MCAR)

Der fallweise Ausschluss ist meist eine sehr naheliegende und unkomplizierte Möglichkeit, mit fehlenden Daten umzugehen. Hier werden für die Analyse nur diejenigen Fälle verwendet, für die alle Variablen beobachtete Werte aufweisen. Im vorliegenden Fall wären das also ausschließlich diejenigen Personen, welche der Verknüpfung ihrer Daten zugestimmt haben. Dabei wird allerdings – meist implizit – davon ausgegangen, dass die Teilstichprobe der Zustimmenden eine einfache Zufallsstichprobe aus allen Befragten darstellt und damit beim Datenausfall die missing completely at random – Annahme gerechtfertigt ist. Ist diese Annahme tatsächlich erfüllt, können auf Basis der verfügbaren Fälle unverzerrte Schätzungen vorgenommen werden, wenn auch die Schätzung wegen geringerer Fallzahl an Effizienz verliert. Ist dies nicht der Fall, führt etwa eine Regressionsanalyse zu verzerrten Parameterschätzungen.

2.2 Multiple Imputation: Missing at random (MAR)

Unter Multipler Imputation (MI; vgl. Rubin 1976, 1987; Weins 2006) versteht man ein Verfahren, bei dem die fehlenden Werte in den Daten mit $m > 1$ plausiblen Werten ersetzt werden, wodurch ebenso viele Datensätze entstehen. Dabei ist es

grundsätzlich irrelevant, ob zu den ausfallbehafteten Variablen auch die abhängige Variable Y gehört. Daher können im Folgenden Y und $\mathbf{X}$ zu $\mathbf{Z}$ zusammengefasst werden, wobei $\mathbf{Z}^{mis} = (Y^{mis}, \mathbf{X}^{mis})$ und $\mathbf{Z}^{obs} = (Y^{obs}, \mathbf{X}^{obs})$.

Das Verfahren der Multiplen Imputation setzt voraus, dass die Beziehung zwischen Datenausfall und der ausfallbehafteten Variable bzw. den ausfallbehafteten Variablen vollständig von beobachteten Daten abhängt (MAR). Das Vorgehen bei der Multiplen Imputation kann in drei Teilschritte zerlegt werden: Imputation, Datenanalyse und Kombination der Ergebnisse.

Im *Imputationsschritt* werden zunächst mit m >1 mehrere plausible Werte für die fehlenden Werte erzeugt. Die MAR-Annahme garantiert dabei, dass $\left(\mathbf{Z}^{\{1\}}, \mathbf{Z}^{\{2\}}, ..., \mathbf{Z}^{\{m\}}\right)$ ergänzte Datensätze aus der Verteilung $f(\mathbf{Z}^{mis} \mid \mathbf{Z}^{obs})$ erzeugt werden können, da nach der Konditionierung auf $\mathbf{Z}^{obs}$ der Datenausfall – im Bezug auf die betrachteten Variablen – zufällig erfolgt.[3] Unabhängig davon wie die Imputationen konkret erzeugt werden, erfolgt im nächsten Schritt die *Datenanalyse* stets in den m generierten ergänzten Datensätzen, wobei Standardverfahren der statistischen Analyse verwendet werden können, z. B. Regressionsanalysen. Im *Kombinationsschritt* werden dann die Ergebnisse der m separaten Analysen aus den verschiedenen imputierten Datensätzen gemäß einfacher Kombinationsregeln (Rubin 1987) miteinander verknüpft. Die MI-Schätzung des Regressionsparameters β erfolgt etwa als einfacher Mittelwert aus den Regressionsparametern β_m, die man aus den $m = 1,...,M$ imputierten Datensätzen erhält: Es sei $m = 1,...,M$ die Zahl der Imputationen, dann ist der MI-Schätzer für den Regressionskoeffizienten β der einfache Durchschnitt über alle M Imputationen: $\overline{\beta}_M = \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}_m$.

Zur Bestimmung der Standardfehler wird zunächst die Varianz innerhalb der imputierten Datensätze (within imputation) $\overline{W}_M = \frac{1}{M}\sum_{m=1}^{M} W_m$ berechnet, mit $W_m = \hat{Var}(\hat{\beta}_m)$ in der m-ten Imputation, sowie die Varianz zwischen den Datensätzen (between imputations) $B_M = \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\beta}_m - \overline{\beta}_M\right)^2$. Der Schätzer für die Ge-

3 Es existiert eine Vielzahl von Varianten zur Erzeugung der Imputationen: z. B. Propensity Score Methoden, Predictive Mean Matching, Diskriminanzanalysen oder logistische Regressionen (vgl. Horton/Kleinman 2007). Bei komplexeren Ausfallmustern bieten sich meist Markov Chain Monte Carlo (MCMC) Methoden an. Hier wird eine Markov-Kette erzeugt, um Ziehungen aus der sogenannten Posteriorverteilung $f(\mathbf{Z}^{\text{mis}} \mid \mathbf{Z}^{\text{obs}})$ zu simulieren. Eine Markov-Kette ist eine Sequenz von Zufallsvariablen, in der die Verteilung eines jeden Elementes vom Wert des vorherigen abhängt. Bei der MCMC-Methode wird eine Kette erzeugt, die lang genug ist, damit die Elemente zu einer stabilen (stationären) Verteilung konvergieren, hier $f(\mathbf{Z}^{\text{mis}} \mid \mathbf{Z}^{\text{obs}})$. Die Implementation der MCMC-Methode kann etwa über den IP-Algorithmus erfolgen (Schafer 1999a).

samtvarianz kombiniert beide Werte auf folgende Weise: $V_M = \overline{W}_M + \frac{M+1}{M} B_M$, wobei $\sqrt{V_M}$ schließlich der Schätzer für die Standardfehler der Regressionskoeffizienten ist.

2.3 Heckman-Korrektur: Missing not at random (MNAR)

Kann man nicht davon ausgehen, dass alle relevanten Einflüsse auf die Zustimmung in den beobachteten Daten erfasst sind, ist eine Selektionskorrektur nach Heckman (zum Folgenden Heckman 1979; Engelhardt 1999) eine mögliche Alternative zur Multiplen Imputation. Der typische Fall ist, dass lediglich in der abhängigen Variable Y fehlende Werte auftreten, während die Kontrollvariablen vollständig beobachtet werden. Das Verfahren ist aber grundsätzlich auch bei Ausfällen in der abhängigen und/oder in mehreren unabhängigen Variablen anwendbar.

Aus der MNAR-Annahme ergibt sich, dass eine Kleinste-Quadrate-Schätzung der Analysegleichung (wiederum wird im Folgenden der Personenindex i zur besseren Lesbarkeit weggelassen)

$$Y = \mathbf{X}\boldsymbol{\beta} + \varepsilon \tag{4}$$

zu verzerrten Parameterschätzungen führt. Um dies zu vermeiden, wird eine zweite Gleichung formuliert, die den Selektionsprozess beschreibt, soweit dieser durch die vorhandenen Daten abzubilden ist.

$$D^* = \mathbf{C}\boldsymbol{\alpha} + \nu \tag{5}$$

Dabei ist D^* eine latente Variable, etwa die latente Bereitschaft, der Datenverknüpfung zuzustimmen. Übersteigt die latente Variable einen bestimmten Wert (z. B. 0), dann stimmt eine Person dem Zusammenspielen zu und sonst nicht:

$$D = \begin{cases} 1 \text{ falls } D^* > 0, \\ 0 \text{ sonst} \end{cases}$$

Demnach wird Y nur für solche Personen beobachtet, für die $\nu > -\mathbf{C}\boldsymbol{\alpha}$ ist, weshalb der Erwartungswert von Y in der Teilpopulation nicht $\mathbf{X}\boldsymbol{\beta}$ ist, sondern $E(Y \mid D = 1, \mathbf{X}) = \mathbf{X}\boldsymbol{\beta} + E(\varepsilon \mid D = 1) = \mathbf{X}\boldsymbol{\beta} + E(\varepsilon \mid \nu > -\mathbf{C}\boldsymbol{\alpha})$.

Heckman (1979) zeigt, dass die Schätzung des Erwartungswertes von Y für die gesamte Population mit dem Problem beim Fehlen einer Variablen vergleichbar ist und ähnliche Konsequenzen für die unverzerrte Schätzung der Regressionspara-

meter β hat. Diese „fehlende Variable" ist hier $E(\varepsilon \mid \nu > -\mathbf{C\alpha})$. Unter der Annahme, dass die Störgrößen in den Gleichungen (4) und (5) einerseits bivariat normalverteilt sind $(\varepsilon, \nu) \sim N(0,0,\sigma_\varepsilon^2, \sigma_\nu^2, \rho_{\varepsilon\nu})$, mit σ für die Varianz der jeweiligen Störgrößen und $\rho_{\varepsilon\nu}$ als ihr Korrelationskoeffizient, und andererseits unabhängig von $\mathbf{X}$ und $\mathbf{C}$, lässt sich diese „fehlende Variable" approximieren durch $E(\varepsilon \mid \nu > -\mathbf{C\alpha}) = \rho_{\varepsilon\nu}\sigma_\varepsilon \lambda$ (vgl. Heckman 1979). Die Variable $\lambda = \frac{\phi(\mathbf{C\alpha})}{\Phi(\mathbf{C\alpha})}$ wird meist als inverse Mills Ratio bezeichnet.

Im sogenannten Two-Step-Verfahren, bei dem die Selektionskorrektur in eine Kleinste-Quadrate-Schätzung einsetzt wird, wird λ in einem ersten Schritt aus der vorhandenen Stichprobe mit $\frac{\phi(\mathbf{C\hat{\alpha}})}{\Phi(\mathbf{C\hat{\alpha}})}$ geschätzt. Die Werte $\mathbf{C\hat{\alpha}}$ sind aus einer Probitregression der Selektionsgleichung (5) zu schätzen und Φ bzw. ϕ bezeichnen die (kumulative) Standardnormalverteilung.

Im zweiten Schritt wird λ in die Analysegleichung eingesetzt. Eine KQ-Schätzung der resultierenden Regressionsgleichung $Y = \mathbf{X\beta} + \rho_{\varepsilon\nu}\sigma_\varepsilon \hat{\lambda} + \varepsilon$ liefert dann eine unverzerrte Schätzung der Regressionsparameter $\boldsymbol{\beta}$, wobei $\rho_{\varepsilon\nu}\sigma_\varepsilon$ der Koeffizient der Variable $\hat{\lambda}$ ist.

Um Kollinearitätsprobleme zu vermeiden sollte dabei $\mathbf{C}$ mindestens eine Instrumentvariable enthalten, also ein Element das zwar in der Selektionsgleichung signifikant ist, nicht jedoch in der Analysegleichung für Y und daher nicht auch in $\mathbf{X}$ enthalten ist (exclusion restriction, vgl. Puhani 2000).

Anstatt des Einsetzens in die KQ-Schätzung kann die Selektionskorrektur allerdings auch durch eine simultane Schätzung beider Gleichungen als Maximum Likelihood (ML)-Schätzung erfolgen, im Folgenden auch als Selektionskorrektur in der ML-Variante bezeichnet. Diese gilt jedoch als noch weniger robust gegenüber Verletzungen der Verfahrensannahmen als die KQ-Variante, wenn auch gesicherte Erkenntnisse hierzu kaum vorliegen (vgl. Winship/Mare 1992). Allerdings ist eine ML-Schätzung für den Fall, dass die abhängige Variable der Analysegleichung eine binäre Variable ist, die einzig möglich Variante der Selektionskorrektur nach Heckman.

3 Empirischer Vergleich

Im Folgenden wird zunächst die Datenbasis, der Ausfallprozess und das Untersuchungsdesign zum Vergleich der Missing Data Verfahren vorgestellt. Dabei wird auf konkrete Forschungsfragen Bezug genommen, die jedoch selbst inhaltlich nicht

von Interesse sind, sondern nur dem Verfahrensvergleich dienen. Danach werden kurz die konkreten Varianten der verwendeten Missing Data Verfahren vorgestellt und schließlich die Ergebnisse des Vergleichs präsentiert.

3.1 Datenbasis und empirischer Ausfallprozess

Die Datenbasis für den empirischen Vergleich bildet eine Befragung zur Kombilohnförderung „Mainzer Modell". Im Rahmen der Evaluation dieser zunächst regional begrenzten, später bundesweit eingesetzten Kombilohnförderung, wurden von TNS Infratest Sozialforschung im Auftrag des Instituts für Arbeitsmarkt- und Berufsforschung (IAB) der Bundesagentur für Arbeit (BA) Geförderte und eine Gruppe ungeförderter Vergleichspersonen befragt. Die Stichprobe umfasste Personen, die im Zeitraum Januar 2001 bis März 2003 von Arbeitslosigkeit in Beschäftigung übergingen (für weitere Details siehe Hartman 2004). Alle Befragten wurden um die Erlaubnis gebeten, ihre Befragungsdaten mit Prozessdaten verknüpfen zu dürfen. Insgesamt stimmten 74,4 % diesem Anliegen zu. Damit ist die Zustimmungsquote, verglichen mit anderen Erhebungen im Bereich der Arbeitsmarktforschung, relativ niedrig.

Grundsätzlich stand der Zugang zur Förderung nach dem Mainzer Modell allen Personen offen, die im Inland eine Beschäftigung aufnehmen dürfen. Daher war die Befragung nicht auf eine bestimmte Teilgruppe beschränkt. Allerdings handelt es sich bei den Geförderten zum großen Teil um Personen im Niedriglohnbereich. Im Anschluss an vorherige Analysen (Kaltenborn et al. 2005; Krug 2009) erfolgt hier zusätzlich eine Beschränkung auf die zwei Gruppen: die Kombilohnbezieher, welche keine abgeschlossene Berufsausbildung aufweisen oder vorher langzeitarbeitslos (mindestens seit einem Jahr ununterbrochen arbeitslos gemeldet) waren und die Vergleichspersonen, für die dies ebenfalls zutrifft.

Für die Umsetzung der Korrekturverfahren (siehe Abschnitt 3.3) ist es relevant, welche Faktoren die Zustimmung zur Zusammenspielung und damit den potentiellen Ausfall von Prozessdaten beeinflussen. Da eine Auseinandersetzung mit dem hier behandelten Ausfallmechanismus bereits an anderer Stelle erfolgte, soll dies hier nur kurz geschehen.[4] Hartmann und Krug (2009) unterschieden in Anlehnung an theoretische Überlegungen zur Teilnahmeverweigerung bei Befragungen zwischen zwei potentiell wirksamen Einflussfaktoren: zum einen allgemeine Einflüsse und zum andern untersuchungsspezifische Einflüsse auf das Zustim-

4 Für die ausführliche Diskussion der Determinanten des Datenausfalls siehe Hartmann/Krug (2009: 125ff.).

mungsverhalten. Als allgemeine, also unabhängig von Thema und Auftraggebern der Untersuchung wirksamen Einflüsse, werden neben den soziodemographischen Merkmalen Alter, Bildung, Region (Ost-/Westdeutschland), Geschlecht und Nationalität auch die subjektive Wertschätzung der Freizeit, die Arbeitszeit, der Haushaltskontext (ist ein Partner vorhanden, leben minderjährige Kinder im Haushalt) sowie das Erwerbseinkommen bzw. fehlende Angaben beim Erwerbseinkommen identifiziert, letzteres als Indikator für ein geringes Vertrauen der Befragten gegenüber dem Interviewer oder der Befragung an sich. Im Gegensatz dazu sind untersuchungsspezifische Einflüsse solche, deren Effekt auf das Zustimmungsverhalten sich speziell aus Gegenstand und Auftraggeber der Befragung ergibt. Im Fall der vorliegenden Befragung ergeben sich aus theoretischer Sicht mögliche Einflüsse einer Tätigkeit im öffentlichen Dienst, der Arbeitslosigkeit vor Beschäftigungsantritt und der Ausbildung. Ebenfalls relevant könnte sein, ob die angetretene Stelle vom (damaligen) Arbeitsamt vermittelt wurde, ob eine aufgenommene Stelle aktuell noch andauert, ob es sich um eine kombilohngeförderte Stelle handelte und ob vor dem Antritt der Beschäftigung Sozialhilfe bezogen wurde.[5]

Tabelle 1, vollständiges Modell, zeigt den empirischen Einfluss der aufgeführten potentiellen Determinanten.[6] Während sich einige der Variablen und Indikatoren als signifikant erweisen, ist eine Vielzahl der Determinanten des allgemeinen Befragungsteilnahmeverhaltens für die Zustimmung zum Datenzusammenspielen scheinbar wenig relevant. Das kann, wie auch die recht hohe Zustimmungsquote, daran liegen, dass sich die Befragten bereits für die Teilnahme an der Befragung entschieden haben und damit bereits eine positive Einstellung vorliegt. Entsprechend liegt der Wert des Pseudo R^2 lediglich bei ca. 0,03, das heißt die aufgeführten Variablen liefern einen nur sehr geringen Erklärungsbeitrag für den Datenausfall. Trotzdem gibt es auch signifikante Unterschiede zwischen zustimmenden und nicht zustimmenden Befragten.

Die nicht signifikanten Variablen können ohne wesentlichen Informationsverlust aus dem Modell entfernt werden (die Informationskriterien AIC bzw. BIC sinken dementsprechend) und es verbleiben nur die empirisch relevanten Faktoren des Ausfallmechanismus (restringiertes Modell).

5 Im Unterschied zu den Analysen in Hartmann/Krug 2009 werden hier nur erwerbstätige Personen untersucht, weshalb sich die verwendeten Variablen sowie die Koeffizienten leicht unterscheiden. Zudem sind die Interviewermerkmale in den verwendeten Daten nicht enthalten.

6 Dazu ist zu beachten, dass die Analyse mit den vollständigen Daten durchgeführt wird; je nach Ausfallszenario wäre der Einfluss mancher Variablen auf das Zustimmungsverhalten für den Anwender also nicht testbar. So kann unter Szenario 4 (vgl. weiter unten) etwa der Einfluss der Region Ostdeutschland auf das Zustimmungsverhalten nicht mehr überprüft werden, da die Variable nur für die Zustimmenden vorliegt.

Es zeigt sich, dass ein verknüpfungsbedingter Datenausfall bei Frauen signifikant häufiger vorkommt, ebenso bei Personen mit ausländischer Staatsbürgerschaft. Dagegen kommt er bei ostdeutschen Befragten seltener vor. Tendenziell stimmen Personen mit hohem Erwerbseinkommen seltener der Verknüpfung zu als Andere. Bei Personen mit Kindern im Haushalt kommt ein verknüpfungsbedingter Datenausfall seltener vor, der vorherige Sozialhilfebezug wirkt in eine ähnliche Richtung. Schließlich stimmen Personen, die ihre Stelle durch Eigeninitiative statt über die Arbeitsvermittlung gefunden haben, ebenfalls seltener der Verknüpfung zu und sind damit häufiger vom Datenausfall betroffen. Demnach erweist sich der Ausfallprozess zwar als systematisch, da hinsichtlich einiger Variablen Unterschiede zwischen Personen mit und ohne Zustimmung bestehen. Die Systematik ist jedoch sehr gering, was am niedrigen R^2 abzulesen ist.

Wie verhält es sich nun zu den in Abschnitt 2 aufgeführten Annahmen über den Ausfallmechanismus? Gegeben, es handelt sich (im restringierten Modell) um eine vollständige Erfassung aller Determinanten des Datenausfalls, hängt die Gültigkeit der Annahmen bezüglich des Ausfallmechanismus *für ein konkretes* Analysemodell nun davon ab, ob und wie stark die Variablen in der Analysegleichung mit den hier aufgeführten Determinanten zusammenhängen. Besteht kein Zusammenhang der Analysevariablen zu den Variablen Geschlecht, Staatsbürgerschaft, etc., so spricht dies für MCAR. Besteht ein Zusammenhang, wird dieser aber mutmaßlich vollständig von den beobachteten Variablen abgedeckt, so spricht dies eher für MAR. Besteht ein Zusammenhang, man vermutet aber, dass es sich nicht um eine vollständige Auflistung der zentralen Determinanten handelt und dass die beobachteten Variablen nur einen Teil des Zusammenhanges erfassen, wäre von MNAR auszugehen. In einem gegebenen Analysefall ist empirisch nur die MCAR-Annahme zu widerlegen, sowohl MAR als auch MNAR sind empirisch nicht testbar (Schafer/Graham 2002: 151).[7]

7 Auch geringe Abweichungen von der MAR haben nicht unbedingt ein Versagen von Verfahren zur Folge, welche diese Annahme treffen. Daher ist selbst mit dem vorliegenden Untersuchungsdesign, bei dem das korrekte Ergebnis bei vollständigen Daten bekannt ist, das Zutreffen der MAR Annahme nicht eindeutig zu belegen (vgl. Schafer/Graham 2002: 151ff.), z. B. durch den Umkehrschluss „wenn die Korrektur durch Multiple Imputation erfolgreich war, dann lag MAR vor".

Tabelle 1 Empirische Determinanten des Datenausfalls (Probitregression mit vollständigen Daten)

Abhängige: Befragter stimmt Zusammenspielen nicht zu (=fehlende Prozessdaten)	vollständiges Modell Koeff. (Std.Fehler)	restringiertes Modell Koeff. (Std.Fehler)
Kombilohnförderung: ja	-0,068 (0,074)	
Alter (Ref.: Bis unter 25 Jahre)		
25 bis 34 Jahre	0,070 (0,104)	
35 bis 44 Jahre	-0,034 (0,103)	
45 bis 54 Jahre	0,018 (0,109)	
55 Jahre oder mehr	-0,191 (0,180)	
Bildung (Ref.: kein Abschluss)		
Volks-/Hauptschulabschluss	-0,149 (0,156)	
Volks-, Hauptschule	-0,084 (0,164)	
Mittlere Reife, POS 10.Klasse	-0,041 (0,154)	
Fachhochschulreife, Abitur	0,078 (0,162)	
Keine Angabe	0,037 (0,226)	
Geschlecht: weiblich	0,125* (0,069)	0,111* (0,063)
Keine deutsche Staatsbürgerschaft	0,221** (0,105)	0,238** (0,102)
Region: Ost	-0,182*** (0,070)	-0,160*** (0,058)
Freizeit sehr wichtig	-0,052 (0,064)	
Teilzeit (< 30 Std.): ja	-0,029 (0,066)	
Bruttoeinkommen: (Ref.: 1.024 Euro oder mehr)		
1 bis 325 Euro	-0,462*** (0,175)	-0,439** (0,171)
326 bis 511 Euro	-0,088 (0,111)	-0,113 (0,109)
512 bis 1.023 Euro	-0,245*** (0,069)	-0,258*** (0,068)
Keine (valide) Einkommensangabe	-0,393*** (0,080)	-0,368*** (0,078)

Abhängige: Befragter stimmt Zusammenspielen nicht zu (=fehlende Prozessdaten)	vollständiges Modell Koeff. (Std.Fehler)	restringiertes Modell Koeff. (Std.Fehler)
Ursprüngliche Beschäftigung zum Befragungszeitpunkt beendet	0,057 (0,059)	
Lebensform: mit Partner	-0,043 (0,061)	
Kinder im Haushalt: ja	-0,204*** (0,068)	-0,216*** (0,057)
Beschäftigt im öffentlichen Dienst: ja	-0,011 (0,094)	
Kumulierte Dauer der Arbeitslosigkeit (Ref.: unter einem Monat)		
Ein bis unter sechs Monate	0,060 (0,080)	
Sechs bis unter zwölf Monate	0,005 (0,090)	
Zwölf bis unter vierundzwanzig Monate	0,146 (0,092)	
Vierundzwanzig Monate und mehr	0,059 (0,112)	
Vor Beschäftigung Sozialhilfe: ja	-0,184** (0,080)	-0,199** (0,079)
Wie wurde die angetretene Stelle gefunden (Ref.: Arbeitsamt)		
Bekannte/Freunde	-0,017 (0,079)	-0,020 (0,078)
Eigene Initiative	0,142* (0,075)	0,142* (0,074)
Sonstige Suchwege	0,059 (0,081)	0,059 (0,080)
Konstante	-0,306 (0,191)	-0,366*** (0,092)
Pseudo R^2	0,029	0,024
p-Wert (chi^2-Test)	0,000	0,000
N	2604	2604
AIC	2934,02	2912,468
BIC	3121,694	2988,711
Likelihood-Ratio-Test auf gemeinsame Insignifikanz der im restringierten Modell ausgelassenen Variablen	chi^2 = 16,45; df = 18; p = 0,63	

**p<0,1; **p<0,05; ***p<0,01*

3.2 Untersuchungsdesign

Für den nachfolgenden empirischen Vergleich wird zunächst auf Basis der vollständigen, unbearbeiteten Befragungsdaten eine Analysegleichung zur Bestimmung des Einflusses einiger unabhängiger Variablen $\boldsymbol{X} = (T,U,\boldsymbol{V}_j,\boldsymbol{W}_k)$ auf die abhängige Variable Y in Form einer Regression aufgestellt:

$$Y = \alpha_0 + \beta\ T + \gamma\ U + \sum_{j=1}^{J} \delta_j V_j + \sum_{k=1}^{K} \eta_k W_k + \varepsilon \qquad (6)$$

Dabei soll T die Variable sein, deren Effekt β auf Y hauptsächlich interessiert, während die anderen Elemente von $\boldsymbol{X}$ lediglich als Kontrollvariablen dienen. Die Ergebnisse aus der Regression mit den vollständigen Befragungsdaten stellen die „wahren Werte"[8] der Koeffizienten dar. Sie dienen somit als Maßstab für den empirischen Vergleich der Missing Data Verfahren, an dem deren Erfolg bei der Korrektur gemessen wird.

Um die Situation nach einer tatsächlichen Datenverknüpfung zu simulieren (die faktisch jedoch nicht stattfand) werden nun einzelne Variablen zu Prozessdaten „umdeklariert"[9], indem bei den Personen ohne Zustimmung die entsprechenden Werte gelöscht werden. Im Anschluss an die Löschung werden die Missing Data Verfahren auf die Daten mit diesen pseudo-fehlenden Werten angewendet und es wird geprüft, ob die erzielten Ergebnisse den „wahren Werten" entsprechen. Dabei werden verschiedene Szenarien durchgespielt (Tabelle 2). Diese unterscheiden sich zunächst darin, ob die Prozessdaten als Kontrollvariablen oder als abhängige Variable in der Analyse dienen. Dienen sie als Kontrollvariablen, so werden schrittweise immer mehr Kontrollvariablen des Analysemodells als Prozessdaten behandelt und mit zustimmungsbedingt fehlenden Werten simuliert. Je mehr Variablen fehlende Werte aufweisen, desto schwieriger sollte die Situation für die Korrekturverfahren werden: Bei der Heckman-Korrektur muss argumentiert werden, dass die Störgröße bzw. die „fehlende(n) Variable(n)" nicht mit den im Modell enthaltenen Variablen $\mathbf{X}$ in Gleichung 4 bzw. $\mathbf{C}$ in Gleichung 5 korreliert sind. Dies wird mit der zunehmenden Zahl ausfallbehafteter Variablen schwieriger. Bei der Multiplen Imputation muss entsprechend für jede einzelne ausfallbehaftete Variable plausibel sein, dass

8 „Wahre Werte" steht hier in Anführungszeichen, da es sich ja auch bei den Ergebnissen mit vollständigen Daten um eine Schätzung handelt.

9 Unabhängig von der Zustimmung lagen bei allen Befragten einige wenige Informationen aus den administrativen Daten der Bundesagentur für Arbeit bereits vor (vgl. Hartmann et al. 2002: 173ff.). Dazu gehört neben der Variablen *Arbeitslosigkeitsdauer* auch die Variable *Stellung im Beruf*. Alle anderen Variablen stammen aus der Befragung, werden aber zum Zweck des Verfahrensvergleiches je nach Szenario wie Prozessdaten mit pseudo-fehlenden Werten versehen.

die MAR-Annahme gegeben ist und auch hier gilt, je mehr solche Variablen, desto eher können Abweichungen von der MAR-Annahme auftreten.

In Szenario 1 wird angenommen, dass nur eine Variable eine ausfallbehaftete Prozessdatenvariable darstellt, nämlich die abhängige Variable Y des Analysemodells (vgl. Gleichung 6). Hierzu werden im Datensatz die Werte der Variable Y für diejenigen Befragten auf „fehlend" gesetzt, welche ihre Zustimmung zu einer Datenverknüpfung verweigert hatten. Auf Basis dieses Datensatzes mit pseudofehlenden Prozessdaten bei den Personen, die einer Zusammenspielung nicht zugestimmt hatten, werden schließlich Schätzungen der Analysegleichung (6) unter Verwendung der in Abschnitt 2 vorgestellten Möglichkeiten des Umgangs mit den fehlenden Werten durchgeführt. Die Ergebnisse der korrigierten Schätzungen können dann mit den tatsächlich auf Basis der vollständigen Daten durchgeführten Schätzungen verglichen werden.

In Szenario 2 wird demgegenüber angenommen, dass lediglich die unabhängige Variable U eine Prozessdatenvariable darstellt und damit nur dort fehlende Werte aufgrund der Verknüpfung entstehen. In Szenario 3 werden Ausfälle in der Variable U und einer Reihe weiterer Kontrollvariablen ($\boldsymbol{V}_j$) simuliert und in Szenario 4 werden schließlich alle oben aufgeführten Kontrollvariablen ($U,\boldsymbol{V}_j,\boldsymbol{W}_k$) auf „fehlend" gesetzt, wenn die Zustimmung zur Datenverknüpfung nicht vorliegt.[10] Die Variable T enthält in keinem der Szenarien fehlende Werte, ist also immer eine Variable aus den Befragungsdaten. Dies soll einerseits garantieren, dass eventuelle Abweichungen zwischen vollständigen und ausfallkorrigierten Daten auf die Qualität der Korrektur der *anderen* Variablen zurückgehen. Andererseits bleiben die Koeffizienten dann über die verschiedenen Szenarien vergleichbar, da sie – z. B. im Fall der Multiplen Imputation – nie imputierte Werte aufweisen.[11]

Tabelle 2 Szenarien des Datenausfalls durch zustimmungspflichtige Datenverknüpfung

Szenario 1	Fehlende Werte in der abhängigen Variable Y
Szenario 2	Fehlende Werte in U, einer unabhängigen Variable
Szenario 3	Fehlende Werte in mehreren unabhängigen Variablen ($U,\boldsymbol{V}_j$)
Szenario 4	Fehlende Werte in allen unabhängigen Variablen ($U,\boldsymbol{V}_j,\boldsymbol{W}_k$)

10 Natürlich wären noch weitere Szenarien bzw. Abstufungen zwischen Szenario 2 und 4 möglich, die vorliegende Analyse beschränkt sich auf diese eine, zumal der Informationsgehalt zusätzlicher Abstufungen wohl eher gering ist.

11 Für Anhaltspunkte, wie die Korrekturverfahren abschneiden, wenn eine *Prozessdatenvariable* im inhaltlichen Interesse der Analyse steht, siehe die Tabellen zum Aufsatz (http://www.gesis.org/forschung-lehre/gesis-publikationen/zeitschriften/mda/jg-4-2010-heft-1/), vor allem die Variable „Arbeitslosigkeitsdauer".

Alle Szenarien werden für zwei unterschiedliche Analysemodelle mit je unterschiedlich starker Beziehung zu den Ausfalldeterminanten durchgespielt.[12] Im Analysemodell 1 wird eine empirische Analyse durchgeführt, die nur gering durch die Selektivität des Datenausfalls betroffen ist. Eine solche Analyse stellt die folgende Untersuchung des Einflusses der Kombilohnförderung „Mainzer Modell" auf die Arbeitszeit dar (vgl. Hartmann/Krug 2009) [13]:

Analysemodell 1:

$$AZEIT = \alpha_0 + \beta\ MZM + \sum_{l=1}^{4} \gamma_l ALO_l + \sum_{j=1}^{J} \delta_j KONT_j^1 + \sum_{k=1}^{K} \eta_k KONT_k^2 + \varepsilon$$

Dabei steht $AZEIT$ für die Arbeitszeit der Person, MZM ist ein binärer Indikator für die Kombilohnförderung im Mainzer Modell (1, falls Mainzer Modell) und ALO steht für Dummyvariablen (1 bis unter 6 Monate, 6 bis unter 12 Monate, 12 bis unter 24 Monate und ab 24 Monate), welche die klassierte „kumulierte Dauer der Arbeitslosigkeit im Erwerbsleben" abbilden. Schließlich stehen $KONT^1$ und $KONT^2$ für zwei Vektoren diskreter sowie kontinuierlicher Kontrollvariablen. Da sich der Datenausfall nur gering auswirkt, sollte Analysemodell 1 eine eher geringe Herausforderung für die Korrekturverfahren darstellen und sie sollten zu Ergebnissen führen, die sehr nah an den „wahren" Werten liegen. Allerdings weisen alle Verfahren noch weitere Anwendungsbedingungen auf, so dass nicht automatisch von einer erfolgreichen Korrektur des Datenausfalls auszugehen ist. So können eventuelle Abweichungen von den wahren Daten auf eine Verletzung der verfahrensspezifischen Anwendungsbedingungen schließen lassen bzw. darauf, dass die Verfahren nicht korrekt durchgeführt wurden.

Analysemodell 2 ist analog zu Analysemodell 1 formuliert, wird hingegen aber so konstruiert, dass das Analysebeispiel stark von der verknüpfungsbedingten Selektivität betroffen ist. Ein solcher Fall ist die folgende Regression, bei der eine der zentralen Determinanten des Zustimmungsverhaltens als abhängige Variable der Regression gewählt wird:

12 Um diese beiden unterschiedlichen Analysemodelle zu identifizieren, musste auf das Ergebnis bei einem fallweisen Ausschluss vorgegriffen werden, weshalb bereits vor der Durchführung streng genommen vorhersehbar war, dass der fallweise Ausschluss in Analysemodell 2 nicht funktioniert. Der Erfolg der anderen beiden Verfahren war natürlich dennoch vollkommen offen.

13 Explizites Ziel im Mainzer Modell war neben der Aktivierung Arbeitsloser zur Aufnahme niedrig entlohnter Beschäftigung vor allem auch die Förderung von Teilzeitbeschäftigung (vgl. Kaltenborn et al. 2005).

Analysemodell 2:

$$EKJA = \alpha_0 + \beta\ ENDE + \sum_{l=1}^{4} \gamma_l ALO_l + \sum_{j=1}^{J} \delta_j KONT_j^1 + \sum_{k=1}^{K} \eta_k KONT_k^2 + \varepsilon$$

Hier ist die abhängige Variable $EKJA$ eine binäre Variable dafür, ob eine Person in der Befragung (eine valide) Einkommensangabe gemacht hat (1, falls ja), $ENDE$ ein binärer Indikator, ob das Beschäftigungsverhältnis, auf das sich die Einkommensangabe bezieht, zum Zeitpunkt der Befragung noch andauert (1, falls nein). $KONT^1$ und $KONT^2$ stehen wiederum für zwei Vektoren diskreter sowie kontinuierlicher Kontrollvariablen, die sich allerdings von denen in beiden Analysemodellen unterscheiden.

Die beiden Analysemodelle sollen zwar möglichst realitätsnah sein, es wird aber nicht der Anspruch erhoben, dass diese Analysen auch genau so durchgeführt würden, wenn es um eine inhaltliche Analyse und nicht um einen Methodenvergleich ginge. Vielmehr wurden sie mit Blick auf die methodischen Schwierigkeiten gewählt. So sind in beiden Analysen sowohl signifikante als auch insignifikante Kontrollvariablen enthalten, was für die Korrekturverfahren bedeutet, dass die Möglichkeit besteht, dass bei fehlerhafter Korrektur signifikante Variablen insignifikant werden und umgekehrt.

3.3 Implementation der Missing Data Verfahren

Die technisch am wenigsten aufwändige Variante ist der fallweise Ausschluss. Hier bedarf es keiner gesonderten Schätzverfahren und/oder Software. Es werden lediglich alle Fälle mit fehlenden Werten aus der Analyse entfernt, und auf die verbliebenen Fälle werden die üblichen statistischen Schätzverfahren angewendet.

Zur Multiplen Imputation der pseudo-fehlenden Prozessdaten wurde in der folgenden Analyse die Methode der verketteten Regressionen („Sequential Regression Multivariate Imputation" SRMI) verwendet, die in der Software IVEWare implementiert ist (Raghunathan et al. 2001; Raghunathan/Solenberger/van Hoeweyk 2002). IVEWare bietet im vorliegenden Anwendungsfall den Vorteil, dass hier komplexe Datenstrukturen berücksichtigt werden können, wie sie in Befragungen häufig auftreten. Neben Imputationsroutinen für kontinuierliche Merkmale bietet IVEWare auch solche für Zähldaten, dichotome und kategoriale Variablen, und es können u. a. Filterbedingungen berücksichtigt werden, z. B. dass nur für Personen mit Kindern die Anzahl oder das Alter dieser Kinder imputiert wird.

Wird bei der Selektionskorrektur zwischen Analyse- und Selektionsgleichung unterschieden, so ist bei der Imputation zwischen dem Analysemodell (die

Variablen in der Regression der Arbeitszeit bzw. der Regression der validen Einkommensangabe) und dem Imputationsmodell (für die Imputation verwendete Variablen) zu unterscheiden. Das Imputationsmodell sollte neben den Variablen des Analysemodells zusätzlich die Variablen enthalten, die a) mit der (den) ausfallbelasteten Variable(n) und b) mit dem Ausfall selbst zusammenhängen (Schafer 1999a: 143). Variablen der Kategorie b) sind vor allem diejenigen Variablen aus Tabelle 1, restringiertes Modell, welche nicht bereits im Analysemodell enthalten sind. Variablen der Kategorie a) sind zwar nicht mit dem Datenausfall korreliert, aber mit der ausfallbelasteten Variable. Neben den Variablen des Analysemodells werden daher im Imputationsmodell noch eine Reihe zusätzlicher Variablen aufgenommen (vgl. Tabelle 5 im Anhang). Für eine ausführliche Diskussion, welche Variablen sinnvollerweise in das Imputationsmodell aufgenommen werden sollen und der Vor- und Nachteile verschiedener Strategien siehe Collins/Schafer/Kam (2001).

Zur Imputation im Rahmen der SRMI Methode wird wie folgt vorgegangen (Raghunathan et al. 2001). Seien $P_1, P_2, \ldots, P_q$ die q Prozessdatenvariablen mit fehlenden Werten für Personen, die ihre Zustimmung zur Datenverknüpfung verweigert haben und sei $\mathbf{B}$ der Vektor der Befragungsvariablen, die also vollständige Beobachtungen enthalten. Im ersten Schritt des ersten Durchgangs wird die Variable P_1 auf $\mathbf{B}$ regressiert und auf dieser Basis die fehlenden Werte in P_1 imputiert. Die Art der Regression ist vom Skalenniveau des Merkmals abhängig.[14] Im zweiten Schritt wird die Variable P_2 auf P_1, $\mathbf{B}$ regressiert – wobei P_1 nun sowohl die beobachteten als auch imputierten Werte enthält, im dritten Schritt P_3 auf P_1, P_2, $\mathbf{B}$, etc. Im zweiten Durchgang wird dann die jeweilige abhängige Variable, zum Beispiel P_1, auf alle anderen Variablen, also $P_2, \ldots, P_q$, $\mathbf{B}$ regressiert, und die im ersten Durchgang imputierten Werte in P_1 werden durch die neuen Werte des zweiten Durchgangs ersetzt. Raghunathan/Solenberger/van Hoeweyk (2002:16) geben an, dass nach zehn Durchgängen die Imputation beendet und ein Datensatz mit vollständigen Daten erzeugt werden kann.

Eine Regressionsanalyse der Gleichung 6 auf Basis dieser imputierten Daten würde nicht die Unsicherheit berücksichtigen, die dadurch entsteht, dass die imputierten Prozessdatenwerte Schätzungen darstellen. Dies wird gewährleistet, indem nicht nur eine (Single Imputation), sondern mehrere Ergänzungen (Multiple Imputation) durchgeführt werden. Zudem enthält die Imputation eine stochastische Komponente (vgl. Fn. 14). Im Allgemeinen gelten bei bis zu 50 % fehlender

14 Implementiert sind lineare, logistische und multinomiale logistische Regressionen sowie Poissonregressionen. Dabei wird allerdings zur Prognose von P_1 den jeweiligen Regressionskoeffizienten aus der Regression von P_1 auf $\mathbf{B}$ ein normalverteilter Zufallsterm hinzuaddiert.

Informationen fünf Imputationen und somit auch fünf verschiedene imputierte Datensätze als ausreichend (Schafer 1999b: 7; kritisch dazu Bodner 2008; Graham/Olchowski/Gilreath 2007).

Zur Schätzung der Regressionsparameter durch Heckmans Selektionskorrektur wurde für das Analysemodell 1 der Stata-Befehl *heckman* verwendet, wobei sowohl die Kleinste-Quadrate als auch die Maximum-Likelihood-Variante implementiert ist. Bei der KQ-Variante wird zuächst die Probitregression der Zustimmung zur Datenverknüpfung durchgeführt $P(D=1|\mathbf{C}) = \phi(\mathbf{C}\alpha)$ (vgl. Gleichung 5) und daraus die inverse Mills Ratio berechnet. Hierzu können nur die Variablen ohne fehlende Werte, also die Befragungsvariablen, als Regressoren verwendet werden. Die inverse Mills Ratio wird dann in eine KQ-Regression der Analysegleichung als zusätzlicher Regressor aufgenommen, wodurch die Schätzung der Regressionskoeffizienten auf Basis der Personen, die der Datenverknüpfung zugestimmt haben, unverzerrt erfolgen kann. In der ML-Variante werden die Probitregression der Zustimmung und die lineare Regression der Arbeitszeit simultan geschätzt. Im Analysemodell 2, das sich u. a. durch eine binäre abhängige Variable auszeichnet, wurde der Stata-Befehl *heckprob* verwendet, wobei ausschließlich eine ML-Schätzung möglich ist und dabei zwei Probitregressionen simultan geschätzt werden.

Typischerweise fällt es schwer, geeignete Instrumentvariablen zur Durchführung der Heckman-Korrektur zu finden (Puhani 2000), das heißt in den konkreten Anwendungsfällen eine oder mehrere Variablen, die in der Selektionsregression signifikant sind, bei Aufnahme in die Analysegleichung jedoch nicht. Grundsätzlich kommen genau die Variablen aus dem restringierten Modell in Tabelle 1 in Frage, welche nicht bereits im jeweiligen Analysemodell vorkommen. Für die Analyse der Arbeitszeit (Analysemodell 1) kommen somit lediglich die Variablen „ausländische Staatsbürgerschaft", „vorheriger Sozialhilfebezug" sowie das „Finden der Stelle über ..." in Frage, da die Variable „Region Ostdeutschland" explizit und die Variablen „Geschlecht" und „Kinder vorhanden" implizit in der Variable „Haushaltskontext" enthalten sind. Für das Bruttomonatseinkommen gilt, dass dieses mit der Arbeitszeit korreliert ist, so dass es ebenfalls als Instrument ausscheidet.[15] Für die Regression der vorhandenen Einkommensangabe der Arbeitszeit (Analysemodell 2) sind die verfügbaren Instrumente entsprechend „ausländische Staatsbürgerschaft", „Region Ostdeutschland" und „Finden der Stelle über ...".

15 Die Einkommensvariable wurde nicht als erklärende Variable in Analysemodell 1 aufgenommen, da die Arbeitszeit das Bruttomonatseinkommen beeinflusst und nicht umgekehrt.

3.4 Ergebnisse des empirischen Vergleiches

In diesem Abschnitt soll die Qualität der Ausfallkorrektur durch die verschiedenen Verfahren miteinander verglichen werden. Ein Ergebnis des Abschnitts 3.1 sowie früherer Analysen mit dem vorliegenden Datensatz (vgl. Hartmann/Krug 2009) war, dass es nur hinsichtlich weniger Variablen systematische Unterschiede zwischen Personen mit und ohne Erlaubnis zur Datenzusammenspielung gibt. Dennoch kann die Analyse bestimmter Fragestellungen *im Einzelfall* stärker oder weniger stark mit diesen wenigen Variablen zusammenhängen. Daher erfolgt der Test der Korrekturverfahren zunächst für eine Fragestellung, welche eher gering von der Selektivität des Datenausfalls betroffen ist (Analysemodell 1), wobei die Ergebnisse unter allen vier Szenarien des Ausfalls den Ergebnissen aus vollständigen Daten gegenüber gestellt werden. Danach wird entsprechend mit dem Analysemodell 2 vorgegangen, das wesentlich stärker von der Ausfallselektivität betroffen ist. Wichtig ist hierbei allerdings, dass ein Anwender im Normalfall statistisch nur bedingt testen kann, ob er bei seinen Analysen von starker oder schwacher Selektivität ausgehen muss, so dass die Entscheidung auf Basis fundierter Überlegungen über den Datenausfall erfolgen muss.

3.4.1 Analysemodell 1: schwache Selektivität

Zunächst wird der Vergleich der verschiedenen Missing Data Verfahren für das Analysemodell 1 durchgeführt, der durch eher schwache Selektivität gekennzeichnet ist. Berücksichtigt man nur die Fälle, in denen die Befragungsdaten keine fehlenden Werte aufweisen, so beträgt die Ausfallquote ca. 25 %. Tabelle 3 zeigt die Ergebnisse der verschiedenen Missing Data Verfahren für die verschiedenen Szenarien im Analysemodell 1 und vergleicht sie mit den Ergebnissen der echten, vollständigen Daten. Der Fokus des Vergleichs liegt auf dem Koeffizienten des Dummies für die Kombilohnförderung (β in Gleichung 6). Darüber hinaus liefert grundsätzlich auch der Vergleich der geschätzten Koeffizienten der pseudo-ausfallbehafteten unabhängigen Variable(n) mit den „wahren" Koeffizienten bei vollständigen Daten Anhaltspunkte zur Bewertung der Missing Data Verfahren. Während die vorliegende Analyse lediglich die Korrektur des Koeffizienten einer nicht ausfallbehafteten Variable durch die ausgewählten Verfahren untersucht, kann aus den Tabellen auch abgelesen werden, wie es sich bei der Korrektur ausfallbehafteter Variablen verhält. Aufgrund ihres Umfangs werden die entsprechenden vollständigen Tabellen nur auf der Webseite der mda (http://www.gesis.org/forschung-lehre/gesis-publikationen/zeitschriften/mda/jg-4-2010-heft-1/) veröffentlicht.

Tabelle 3 Analysemodell 1 – Koeffizienten und Standardfehler der Dummyvariable „Kombilohnförderung (MZM)" nach Szenarien

OLS [a]- Regressionen der wöchentlichen Arbeitszeit in Stunden	vollständige Daten Koeff. (Std.Fehler)	fallweiser Ausschluss Koeff. (Std.Fehler)	Multiple Imputation Koeff. (Std.Fehler)	Heckman-Korrektur (OLS) Koeff. (Std.Fehler)	Heckman-Korrektur (ML) Koeff. (Std.Fehler)
Szenario 1	-4,769***	-4,462***	-4,135***	-4,402***	-4,270***
	(0,550)	(0,632)	(0,660)	(0,641)	(0,682)
Szenario 2	-4,769***	-4,462***	-4,752***	-4,370***	-4,180***
	(0,550)	(0,632)	(0,551)	(0,650)	(0,680)
Szenario 3	-4,769***	-4,462***	-4,349***	-4,342***	-4,131***
	(0,550)	(0,632)	(0,674)	(0,663)	(0,664)
Szenario 4	-4,769***	-4,462***	-4,455***	-4,296***	-4,041***
	(0,550)	(0,632)	(0,577)	(0,685)	(0,665)

**p<0,05; **p<0,01; ***p<0,001*
[a] Ausnahme SSM(ML)

In der zweiten Spalte der Tabelle 3 findet sich die Schätzung auf Basis der vollständigen Daten. Die Ergebnisse zeigen, dass die Kombilohnförderung einen signifikant negativen Einfluss auf die Arbeitszeit hatte. Im Mittel arbeiten Geförderte 4,8 Stunden weniger als ungefördert Beschäftigte, dies gilt (natürlich) für alle Szenarien gleichermaßen.

Fallweiser Ausschluss

Die Analyse mit fallweisem Ausschluss kommt, da sie nur schwach von Selektivität betroffen ist, zu sehr ähnlichen Ergebnissen wie eine Analyse unter Vollständigkeit der Daten. Allerdings liegt bei fallweisem Ausschluss eine leichte Unterschätzung des negativen Einflusses der Förderung vor (d. h. der negative Effekt wird als leicht geringer eingeschätzt). Bedingt durch die geringere Fallzahl zeigt sich auch ein größerer Standardfehler. Dennoch bleibt der Effekt auf dem 0,1 %-Niveau signifikant, so dass sich die aus der Analyse ergebende Schlussfolgerung nicht verändert. Auch bei den anderen Variablen (Tabelle 6, siehe http://www.gesis.org/forschung-lehre/gesis-publikationen/zeitschriften/mda/jg-4-2010-heft-1/) halten sich die Unterschiede zu den Ergebnissen bei vollständigen Daten in Grenzen. Lediglich beim Dummy für die kumulierte Arbeitslosigkeitsdauer „Sechs bis unter 12 Monaten" wird aus einem

insignifikanten ein stärkerer und nun signifikanter Effekt.[16] Stets bleiben (nicht) signifikante Einflüsse (nicht) signifikant, wenn es auch zu Unterschieden in der Größe der Koeffizienten kommt, etwa den Dummys für „Angestellte" bzw. „Reinigungsgewerbe". Beim Dummy „Freizeit sehr wichtig" dreht sich zwar das Vorzeichen, der entsprechende Koeffizient ist und bleibt jedoch insignifikant. Aufgrund des Datenausfallmusters bei der Datenverknüpfung (die ausfallbehafteten Beobachtungen sind stets dieselben, unabhängig davon, welche Variablen betrachtet werden) verändert sich die Schätzung über die verschiedenen Szenarien nicht.

Multiple Imputation

Im Gegensatz zum fallweisen Ausschluss unterscheiden sich bei den anderen statistischen Korrekturverfahren die Ergebnisse je nach Szenario. Zunächst zur Multiplen Imputation: In Szenario 1 bestehen lediglich in der abhängigen Variable des Analysemodells fehlende Werte. Hier steigt zum einen der Standardfehler im Vergleich zur Analyse mit vollständigen Daten, zum anderen schneidet die Imputation schlechter ab als der fallweise Ausschluss (oder auch die Heckman-Korrektur in der KQ-Variante, vgl. unten). Kaum Unterschiede ergeben sich zwischen Multipler Imputation und der Selektionskorrektur nach Heckman in der Maximum-Likelihood-Variante. In Szenario 2 mit nur einer pseudo-ausfallbelasteten Kontrollvariable zeigt sich hingegen, dass die Schätzung für den Effekt der Förderung sehr nahe an den „wahren" Ergebnissen mit vollständigen Daten liegt. Enthalten mit Szenario 3 und 4 ein großer Teil bzw. alle unabhängigen Variablen fehlende Werte, so entfernt sich die Schätzung vom Referenzwert aus den vollständigen Daten. Sie ist aber immer noch nahe am Ergebnis mit vollständigen Daten, wenn sie auch erhöhte Standardfehler aufweist.

Heckman-Korrektur

Die Selektionskorrektur in der KQ-Variante weist über alle Szenarien eine ähnliche Qualität der Ergebnisse auf. Dabei sind die Schätzergebnisse stets etwas weiter von den Ergebnissen mit vollständigen Daten entfernt als beim fallweisen Ausschluss. Auch die Standardfehler sind vergleichsweise groß. Im Übergang von Szenario 1 zu Szenario 2 zeigt sich im Gegensatz zur Imputation allerdings eine nur geringe Veränderung der Koeffizienten. Sie entsteht dadurch, dass die Variable ALO in Sze-

16 Im Gegensatz dazu führt beim Dummy „Arbeit sehr wichtig" die leichte Erhöhung des Standardfehlers zur Insignifikanz des Effektes, der Koeffizient war aber bereits bei vollständigen Daten an der Schwelle zur Insignifikanz.

nario 2 nun ausfallbehaftet ist und daher nicht zur Berechnung von λ durch die Selektionsgleichung verwendet werden kann. Im Vergleich zur Multiplen Imputation liefert die KQ-Variante der Heckman-Korrektur also bei fehlenden Werten in der unabhängigen Variable schlechtere, in einer abhängigen Variable bessere Ergebnisse und bei mehreren unabhängigen Variablen in etwa ähnliche (Szenario 3) bis leicht schlechtere Ergebnisse (Szenario 4). Demgegenüber liefert die ML-Variante der Selektionskorrektur nach Heckman die Koeffizientenschätzungen mit der größten Abweichung zum vollständigen Datensatz, am stärksten ist die Abweichung in Szenario 4. Doch selbst hier hält sich die Abweichung in Grenzen. Der Koeffizient ist auch in Szenario 4 signifikant von Null verschieden und mit 4,041 im Vergleich zum „wahren Wert" 4,769 um nur 15 % unterschätzt.

3.4.2 Analysemodell 2: starke Selektivität

In Tabelle 4 sind die Ergebnisse der Korrekturverfahren unter der Bedingung einer starken Betroffenheit des Analysemodells von der Selektivität des Datenausfalls abgebildet. Berücksichtigt man wiederum nur die Fälle, in denen die Befragungsdaten keine fehlenden Werte aufweisen, so ist die Ausfallquote mit ca. 22 % leicht geringer als im Analysemodell 1. Erneut liegt der Schwerpunkt des Vergleichs auf dem Koeffizienten β, das heißt in Analysemodell 2 dem Koeffizienten der Variable „Beschäftigung zum Interviewzeitpunkt bereits beendet". Für die vollständigen Tabellen sei auf Anhang B (siehe http://www.gesis.org/forschung-lehre/gesis-publikationen/zeitschriften/mda/jg-4-2010-heft-1/) verwiesen. In den vollständigen Daten zeigt sich hier, dass die Wahrscheinlichkeit sinkt, dass Befragte eine valide Einkommensangabe machen, wenn die angesprochene Beschäftigung zum Zeitpunkt der Befragung bereits beendet ist.

Fallweiser Ausschluss

Der starke Einfluss der Selektivität zeigt sich deutlich, wenn man die Ergebnisse der Regression mit vollständigen Daten und bei fallweisem Ausschluss vergleicht. Ist der Effekt ohne verknüpfungsbedingten Datenausfall mit -0,174 signifikant negativ, so zeigt sich in der Regression mit fallweisem Ausschluss ein weitaus kleinerer und insignifikanter Effekt von -0,089. Damit kommt es durch den Datenausfall im Vergleich zum Fall mit schwacher Selektivität zu inhaltlich deutlich anderen Aussagen. In ähnlicher Weise sind auch andere Variablen des Analysemodells 2 betroffen. So ist etwa der Effekt des Geschlechts nach fallweisem Ausschluss nicht mehr signifikant, während umgekehrt der Effekt des Alters nun als signifikant für die valide Einkommensangabe erscheint.

Tabelle 4 Analysemodell 2 – Koeffizienten und Standardfehler der Dummyvariable „Beschäftigung zum Befragungszeitpunkt bereits beendet (*ENDE*)" nach Szenarien

Probitregressionen „Einkommensangabe vorhanden"	vollständige Daten Koeff. (Std.Fehler)	fallweiser Ausschluss Koeff. (Std.Fehler)	Multiple Imputation Koeff. (Std.Fehler)	Heckman-Korrektur (ML) Koeff. (Std.Fehler)
Szenario 1	-0,174**	-0,089	-0,085	-0,066
	(0,065)	(0,075)	(0,065)	(0,096)
Szenario 2	-0,174**	-0,089	-0,168**	-0,066
	(0,065)	(0,075)	(0,065)	(0,086)
Szenario 3	-0,174**	-0,089	-0,176**	-0,061
	(0,065)	(0,075)	(0,065)	(0,069)
Szenario 4	-0,174**	-0,089	-0,171*	-0,079
	(0,065)	(0,075)	(0,067)	(0,082)

**p<0,05; **p<0,01; ***p<0,001*

Multiple Imputation

Geht man davon aus, dass die Selektivität durch beobachtete Daten erfasst werden kann, so sollte die Multiple Imputation die bei fallweisem Ausschluss vorliegenden Verzerrungen beheben können. In Szenario 1 mit fehlenden Werten in der abhängigen Variable ist dies nicht der Fall. Der Koeffizient der Variable „Beschäftigung beendet" unterscheidet sich kaum vom verfälschten Wert bei fallweisem Ausschluss.[17] Fehlen die Werte allerdings lediglich in einer der unabhängigen Variablen (Szenario 2), so gelingt es der Multiplen Imputation die ausfallbedingten Verzerrungen weitgehend auszugleichen. Der Koeffizient ist mit -0,168 sehr nahe am Ergebnis mit vollständigen Daten und wird ebenfalls als signifikant ausgewiesen. In Szenario 2 muss lediglich MAR im Bezug auf die Variable ALO vorliegen, in Szenario 3 muss diese Annahme für ALO und zusätzlich alle Variablen $KONT^{l}$ erfüllt sein. Aber auch hier liegt das Ergebnis der Korrektur mit -0,176 sehr nahe am „wahren" Koeffizienten und ist signifikant von Null verschieden. Zeigen mit Szenario 4 alle Kontrollvariablen fehlende Werte, wird der Koeffizient mit -0,171 ebenfalls als signifikant ausgewiesen, wenn auch der Standardfehler im Vergleich zu den

17 Um zu testen, ob sich das Ergebnis der Multiplen Imputation mit mehr als 5 Imputationen verbessert, wurde die Analyse auch mit 10 und 20 Imputationen durchgeführt. Die Koeffizienten und Standardfehler (in Klammern) unterschieden sich mit -0,087 (0,064) nach 10 und -0,084 (0,066) nach 20 Imputationen nicht wesentlich von den Ergebnissen mit 5 Imputationen.

vollständigen Daten etwas ansteigt. Damit führt die Multiple Imputation auch im Fall starker Selektivität zu einer guten Korrektur des Datenausfalls. Dies gilt jedoch nicht für fehlende Werte in der abhängigen Variable.

Heckman-Korrektur

Im Gegensatz zur Multiplen Imputation geht das Heckman-Verfahren auch von unbeobachteten Einflüssen auf den Datenausfall aus und berücksichtigt diese bei der Ausfallkorrektur. Ist die abhängige Y-Variable binär, so kommt ausschließlich die ML-Variante der Heckman-Korrektur in Frage. Trotz der weniger restriktiven Annahmen über den Ausfallprozess zeigt sich das Heckman-Verfahren allerdings als wenig erfolgreich bei der Korrektur des Datenausfalls. In allen vier Szenarien wird das Ausmaß des negativen Effektes der Variable „Beschäftigung beendet" mit Weten zwischen -0,06 und -0,08 im Vergleich zum „wahren Wert" -0,174 deutlich unterschätzt. Zum andern wird der Effekt im Gegensatz zu den Ergebnissen aus vollständigen Daten als insignifikant ausgewiesen. Damit liegt die Heckman-Korrektur sogar noch unter den Ergebnissen des fallweisen Ausschlusses.

4 Zusammenfassung und Diskussion der Ergebnisse

Beschränkt man sich beim Umgang mit fehlenden Prozessdaten auf den fallweisen Ausschluss, so setzt man voraus, dass der Datenausfall ohne Zusammenhang mit den Analysevariablen ist. In der durchgeführten Fallstudie führte eine schwache Selektivität des Datenausfalls im Bezug auf das entsprechende Analysemodell zu nur geringen Unterschieden im Vergleich zu Analysen mit vollständigen Daten. Hat ein Anwender also Grund für die Annahme, dass der Datenausfall bei der Zusammenspielung zufällig oder zumindest nur schwach systematisch erfolgt, so erscheint der fallweise Ausschluss als einfache und sinnvolle Alternative zu komplexen Korrekturverfahren. Lag der Fall vor, dass die Analysevariablen stark mit den Determinanten des Ausfallprozesses zusammenhängen, so verwundert es aufgrund der MCAR-Annahme des fallweisen Ausschlusses kaum, dass gravierende Abweichungen zu den „wahren" Ergebnissen mit vollständigen Daten entstanden.

Die Multiple Imputation setzt als Verfahren die Annahme voraus, dass der Datenausfall zwar systematisch ist, diese Systematik aber durch die beobachteten Daten erfasst wird. Beginnend mit der Situation schwacher Selektivität, lieferte die Multiple Imputation bei fehlenden Werten in einer oder mehreren unabhängigen Variablen entweder leicht bessere oder ähnlich gute Ergebnisse wie der fallweise

Ausschluss. War die abhängige Variable eine Prozessdatenvariable, führte die Verwendung der Multiplen Imputation allerdings zu einer leichten Verschlechterung der Schätzung. Insgesamt waren in allen Szenarien die Abweichungen von Ergebnissen mit vollständigen Daten aber eher gering, so dass sich das Verfahren im Großen und Ganzen ähnlich gut für die Ausfallkorrektur geeignet zeigte, wie der fallweise Ausschluss. Bei der Analyse mit starker Selektivität erwies sich die Imputation als das einzige unter den getesteten Korrekturverfahren, das für den interessierenden Koeffizienten eine Schätzung lieferte, welche das Ergebnis aus den vollständigen Daten reproduzieren konnte. Dies gelang bei fehlenden Werten in einer, mehreren und sogar allen Kontrollvariablen, allerdings aber gerade nicht für die Ausfälle in der Y-Variable, wo sich das Ergebnis bei Imputation kaum von dem beim fallweisen Ausschluss unterschied.

Für die Heckman-Korrektur ist trotz der MNAR-Annahme zu beobachten, dass sie sowohl beim Vorliegen schwacher als auch starker Selektivität keine wesentliche Verbesserung im Vergleich zum fallweisen Ausschluss erzielte. Letzteres bedeutet wiederum, dass dort aufgetretene Verzerrungen nicht ausgeglichen wurden. Dass die Heckman-Korrektur die Schätzung im Vergleich zum fallweisen Ausschluss nicht verbessert, mag darin begründet sein, dass sie zwar die geringsten Annahmen über den Ausfallprozess macht, aber sehr strenge Anforderungen an die Verteilung der Störgrößen in Selektions- und Analysegleichung (bivariate Normalverteilung) stellt sowie auf Instrumente angewiesen ist. Womöglich sind die verwendeten Instrumente im vorliegenden Fall nicht ausreichend und/oder die Verteilungsannahmen sind nicht gerechtfertigt.

Da es sich um eine Fallstudie handelt, können diese Ergebnisse keine uneingeschränkte Verallgemeinerbarkeit beanspruchen. Zum einen geht es hier nur um den Ausfallmechanismus „keine Zustimmung zur Datenverknüpfung" und nicht um einen generellen Vergleich der Leistungsfähigkeit der Missing Data Verfahren. Zum andern kann in anderen Studien, die Prozess- und Befragungsdaten verknüpfen, der Ausfallprozess eventuell besser oder auch schlechter mit den verfügbaren Informationen erfasst werden, als mit den hier verwendeten Daten. Dies kann natürlich die Genauigkeit beeinflussen, mit der es den Missing Data Verfahren gelingt, den verknüpfungsbedingten Datenausfall zu korrigieren. So kann sich die Leistungsfähigkeit der hier untersuchten Verfahren in anderen Forschungszusammenhängen von der hier Berichteten natürlich unterscheiden. Allerdings ist anzunehmen, dass die allgemeinen empirischen Determinanten des Zustimmungsverhaltens zur Datenverknüpfung – und damit des Datenausfalls – auch in anderen Befragungen ähnlich wirksam sind, und damit die Selektivität der Stichprobe nach Zusammenspielung ähnlich sein wird (unabhängig davon, ob Determinanten nun beobachtete

oder unbeobachtete Einflüsse darstellen). Daher wird diese Fallstudie einige wichtige Hinweise für die Anwender liefern können, die in einer konkreten Forschungsarbeit beide Datenquellen nutzen wollen und vor der Wahl eines entsprechenden Korrekturverfahrens stehen.

Schließlich ist noch zu diskutieren, ob die Anwendung der Korrekturverfahren *in dieser Fallstudie* nicht noch in der Hinsicht verbessert werden könnte, so dass sie bessere Ergebnisse liefert. Eindeutig (und negativ) ist die Antwort lediglich beim fallweisen Ausschluss, da hier keine weiteren Varianten möglich sind. Für die Multiple Imputation stellt sich diese Frage lediglich bei fehlenden Werten in der abhängigen Variable. Hier fällt auf, dass das Ergebnis der Korrektur sowohl bei schwacher als auch starker Selektivität zu wünschen übrig lässt. Ob der Grund hierfür in der Verletzung der MAR-Annahme lag, oder ob es sich um ein generelles Problem bei der Imputation fehlender Werte in abhängigen Variablen einer Regression handelt (z. B. Hippel 2007)[18], kann hier nicht abschließend geklärt werden. Für ein besseres Abschneiden der Heckman-Korrektur könnten eventuell mehr oder bessere Instrumentvariablen helfen. Dies setzt allerdings voraus, dass es möglich ist, trotz der vorangegangenen intensiven Auseinandersetzung mit Determinanten des Datenausfalles (Hartmann/Krug 2009) noch weitere Determinanten zu finden, die zudem noch die Instrumenteneigenschaft aufweisen. Zumindest in den vorliegenden Daten ist dies wohl wenig wahrscheinlich. Hinzu kommt, dass laut Wilde (2000) die bivariate Variante des Heckmanverfahrens (Analysemodell 2) in vielen Fällen auch ohne Instrumentvariablen funktionieren sollte. In dem Fall bleibt nur der Schluss, dass in der vorliegenden Analyse die verfahrensbedingten Annahmen über die unbeobachtete Heterogenität (bivariate Normalverteilung, Unkorreliertheit mit den enthaltenen Variablen) nicht zutreffen. Dass diese sehr restriktiv sind und ihre Verletzung problematisch für die Qualität der Heckman-Korrektur ist, dazu vergleiche etwa Winship/Mare (1992), zum Umgang mit Annahmeverletzungen siehe z. B. Newey (2009).

18 Hippel zeigt, dass im Rahmen einer Multiplen Imputation solche Beobachtungen, welche ausschließlich in der Y-Variable fehlende Werte aufweisen, zunächst keinerlei Information zur Ausfallkorrektur beitragen, sondern die Schätzung der Koeffizienten lediglich mit statistischem Rauschen versehen. Zusätzliche Information enthalten die imputierten Werte allerdings – wie in der vorliegenden Analyse – durch die a)- und b)-Variablen (Tabelle 5; vgl. Hippel 2007: 108ff.), so dass das schlechte Abschneiden der Multiplen Imputation in Szenario 1 nur teilweise durch die Ergebnisse von Hippel erklärbar ist.

5 Schluss

Die Verknüpfung von Informationen über ein Individuum aus verschiedenen Quellen (Record Linkage) gewinnt in der empirischen Sozialforschung zunehmend an Bedeutung (z. B. Schnell/Bachteler/Reiher 2009), bedingt unter anderem auch durch die verstärkte Aufbereitung von Prozessdaten für wissenschaftliche Zwecke. Aus Datenschutzgründen sind solche Verknüpfungen, etwa die von Befragungsdaten mit Prozessdaten, oft zustimmungspflichtig, so dass es hier zu Datenausfällen kommen kann. In der vorliegenden Fallstudie wurden die Resultate mehrerer Verfahren beim Umgang mit fehlenden Werten im Rahmen einer zustimmungspflichtigen Datenverknüpfung empirisch miteinander verglichen.

Fasst man die Ergebnisse des Verfahrensvergleiches zusammen, so sprechen diese – unter Berücksichtigung der stets eingeschränkten Verallgemeinerbarkeit solcher Fallstudien – dafür, bei Bedarf eine Anreicherung von Befragungsdaten mit Prozessdaten vorzunehmen, auch wenn es durch die Zustimmungspflichtigkeit zu Datenausfällen kommt. Ein Grund ist, dass nach bisherigen Erkenntnissen der Fall starker Korrelationen von Analysevariablen mit Zustimmungsdeterminanten wohl eher selten ist, da es nur wenige solcher Determinanten zu geben scheint. Daher werden viele Analysen von nur schwacher Selektivität betroffen sein und in diesem Fall sind – bei sachgerechter Anwendung – alle hier behandelten Korrekturverfahren ähnlich gut geeignet. Geht ein Anwender trotz schwacher Selektivität von starkem Ausfall und entweder beobachteten oder zusätzlichen unbeobachteten Ausfalldeterminanten aus (und verwendet entsprechend komplexere Korrekturverfahren), so begeht er laut dieser Fallstudie wohl keinen Fehler.

Anders ist es allerdings, wenn der wohl eher seltenere Fall der starken Selektivität vorliegt. Dies ist etwa der Fall, wenn Aspekte der Datenqualität im Zentrum der Untersuchung stehen, wie etwa im Analysemodell 2. In diesem Fall riskieren Forscher mit der Verwendung des fallweisen Ausschlusses einerseits und der Heckman-Korrektur andererseits – zumindest unter zu der vorliegenden Untersuchung ähnlichen Bedingungen – falsche inhaltliche Schlussfolgerungen aus ihren Analysen. Dies gilt bei der Verwendung von ausfallbelasteten Prozessdatenvariablen als abhängige Regressionsvariablen auch für die Multiple Imputation. Nutzt ein Anwender eine oder mehrere Prozessdatenvariablen als Kontrollvariablen, so scheint die Multiple Imputation ein Verfahren, welches – wiederum unter zu der vorliegenden Untersuchung ähnlichen Bedingungen –, helfen kann, falsche Schlussfolgerungen zu vermeiden. Je weiter natürlich die Bedingungen einer Analyse mit verknüpften Daten von denen in dieser Fallstudie abweichen, etwa hinsichtlich des Analysemodells oder des Themas der Befragung, desto weniger aussagekräftig sind die hier angeführten Ergebnisse.

Literatur

Allmendinger, J. und A. Kohlmann, 2005: Datenverfügbarkeit und Datenzugang am Forschungsdatenzentrum der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- und Berufsforschung. Allgemeines Statistisches Archiv 88: 159-182.

Bodner, T. E., 2008: What Improves With Increased Missing Data Imputations? Structural Equation Modeling 15(4): 651-675.

Collins, L. M., J. L. Schafer und C. M. Kam, 2001: A Comparison of Inclusive and Restrictive Missing-Data Strategies in Modern Missing-Data Procedures. Psychological Methods 6: 330-351.

Engelhardt, H., 1999: Lineare Regression mit Selektion: Möglichkeiten und Grenzen der Heckman-Korrektur. Kölner Zeitschrift für Soziologie und Sozialpsychologie 51: 706-723.

Graham, J. W., A. E. Olchowski und T. D. Gilreath, 2007: How Many Imputations Are Really Needed? – Some Practical Clarifications of Multiple Imputation Theory. Prevention Science 8(3): 206-213.

Graham, J. W., 2009: Missing Data Analysis: Making It Work in the Real World. Annual Review of Psychology 60: 549-576.

Hartmann, J., 2004: Repräsentative Erhebung zur Evaluation des Mainzer Modells. S. 49-139 in: T. Gewiese, J. Hartmann, G. Krug, und H. Rudolph (Hg.): Das Mainzer Modell aus Sicht der Arbeitnehmer und Betriebe. Befunde aus der Begleitforschung. Nürnberg: IAB. http://doku.iab.de/externe/2004/k040823w09.pdf (04.05.2009).

Hartmann, J., K. Brink, R. Jäckle und N. Tschersich, 2008: IAB-Haushaltspanel im Niedrigeinkommensbereich – Methoden- und Feldbericht. FDZ Methodenreport 7/2008 Nürnberg: IAB. http://doku.iab.de/fdz/reporte/2008/MR_07-08.pdf (04.05.2009).

Hartmann, J., A. Holllederer, B. Kaltenborn, H. Rudolph, A. Vanselow, C. Weinkopf, und E. Wiedemann, 2002: Vom arbeitsmarktpolitischen Sonderprogramm CAST zur bundesweiten Erprobung des Mainzer Modells. 2. Zwischenbericht des Forschungsverbunds „Evaluierung CAST". BMWA-Dokumentation Nr. 516.

Hartmann, J. und G. Krug, 2009: Verknüpfung von personenbezogenen Prozess- und Befragungsdaten. Selektivität durch fehlende Zustimmung der Befragten? Zeitschrift für ArbeitsmarktForschung 42(2): 121-139.

Heckman, J. J., 1979: Sample Selection Bias as a Specification Error. Econometrica 47(1): 153-162.

Hippel, P. T., 2007: Regression With Missing Ys: An Improved Strategy for Analyzing Multiply-Imputed Data. Sociological Methodology 37(1): 83-117.

Horton N. J. und K. P. Kleinman, 2007: Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models. The American Statistician 61: 79-90.

Horton, N. J. und S. R. Lipsitz, 2001: Multiple Imputation In Practice: Comparison of Software Packages for Regression Models With Missing Variables. American Statistician 55: 244-254.

Kaltenborn, B., G. Krug, H. Rudolph, C. Weinkopf und E. Wiedemann, 2005: Evaluierung der arbeitsmarktpolitischen Sonderprogramme CAST und Mainzer Modell. Bundesministerium für Wirtschaft und Arbeit. Forschungsbericht Nr. 552, Berlin.

Krug, G., 2009: In-Work Benefits For Low-Wage Jobs. Can Additional Income Reduce Employment Stability? European Sociological Review 25(4): 459-474.

Little, R. J. A. und D. B. Rubin, 1987: Statistical Analysis with Missing Data. New York: Wiley.

Newey, W. K., 2009: Two-Step Series Estimation of Sample Selection Models. Econometrics Journal 12: 217-229.

Puhani, P. A., 2000: The Heckman Correction for Sample Selection and Its Critique. Journal of Economic Surveys 14(1): 53-68.

Raghunathan, T. E., J. M. Lopkowski, J. van Hoeweyk und P. Solenberger, 2001: A Multivariate Technique for Multiply Imputing Missing Values Using a Series of Regression Models. Survey Methodology 27: 85-96.

Raghunathan, T. E., P. Solenberger und J. van Hoeweyk, 2002: IVEware: Imputation and Variance Estimation Software. User Guide. http://www.isr.umich.edu/src/smp/ive (04.05.2009).

Schnell, R., T. Bachteler und J. Reiher, 2009: Entwicklung einer neuen fehlertoleranten Methode bei der Verknüpfung von personenbezogenen Datenbanken unter Gewährleistung des Datenschutzes. MDA – Methoden, Daten, Analysen. Zeitschrift für Empirische Sozialforschung 3(2): 203-217. http://www.gesis.org/fileadmin/upload/forschung/publikationen/zeitschriften/mda/Vol.3_Heft_2/06_Schnell_et_al.pdf (14.4.2010).

Rässler, S., 2002: Statistical Matching. A Frequentist Theory, Practical Applications and Alternative Bayesian Approaches. New York, Berlin: Springer.

Ridder, G., 1992: An Empirical Evaluation of Some Models For Non-Random Attrition In Panel Data. Structural Change and Economic Dynamics 3(2): 337-355.

Rubin, D. B., 1976: Inference With Missing Data (With Discussion). Biometrica 63: 581-592.

Rubin, D. B., 1987: Multiple Imputation for Nonresponse in Surveys. New York: Wiley.

Schafer, J. L., 1999a: Analysis of Incomplete Multivariate Data. London, u. a.: Chapman & Hall/CRC.

Schafer, J. L., 1999b: Multiple Imputation: A Primer. Statistical Methods in Medical Research 8(1): 3-15.

Schafer, J. L. und J. W. Graham, 2002: Missing Data: Our View of the State of the Art. Psychological Methods 7(2): 147-177.

Weins, C., 2006: Multiple Imputation. S. 205-216 in: J. Behnke, V. Gschwend, D. Schindler, und K.-U. Schnapp (Hg.): Methoden der Politikwissenschaft. Neuere quantitative Methoden Analyseverfahren. Baden-Baden: Nomos.

Wilde, J., 2000: Identification of Multiple Equation Probit Models With Endogenous Dummy Regressors. Economics Letters 69: 309-312.

Winship, C. und R. D. Mare, 1992: Models for Sample Selection Bias. Annual Review of Sociology 18: 327-50.

Wirth, H. und W. Müller, 2004: Mikrodaten der amtlichen Statistik als eine Datengrundlage der empirischen Sozialforschung. S. 93-127 in: A. Diekmann (Hg.): Methoden der Sozialforschung (Sonderheft 44 der Kölner Zeitschrift für Soziologie und Sozialpsychologie). Wiesbaden: VS Verlag für Sozialwissenschaften.

Anschrift des Autors

Dr. Gerhard Krug
Institut für Arbeitsmarkt- und Berufsforschung (IAB)
der Bundesagentur für Arbeit (BA)
Weddigenstr. 20-22
90478 Nürnberg
Gerhard.Krug@iab.de

Anhang

Tabelle 5 Zusätzliche Variablen im Imputationsmodell

	Analysefall 1	**Analysefall 2**
Suchanstrengungen (Zahl der Suchwege)	a)-Variable	Analysevariable
Bereits einmal Stelle wegen zu niedrigem Einkommen abgelehnt (Dummy)	a)-Variable	Analysevariable
Wie wurde die Stelle gefunden (Arbeitsvermittlung; Bekannte/Freunde; Eigene Initiative; Sonstiges)	b)-Variable	b)-Variable
Sozialhilfebezug vor Beschäftigungsaufnahme (Dummy)	b)-Variable	Analysevariable
Von ABM in Beschäftigung (Dummy)	a)-Variable	Analysevariable
Von and. Maßnahme in Beschäftigung (Dummy)	a)-Variable	a)-Variable
Vorher Weiterbildung (Dummy)	a)-Variable	Analysevariable
Vorher Lohnersatzleistungen vom Arbeitsamt (Arbeitslosengeld; Arbeitslosenhilfe; Keine)	a)-Variable	Analysevariable
Beschäftigung befristet (Dummy)	a)-Variable	a)-Variable
Bruttolohn in Euro (falls Angabe)	b)-Variable	a)-Variable
Bruttolohn: Angabe nicht verweigert (Dummy)	b)-Variable	Analysevariable
Geschlecht weiblich (Dummy)	b)-Variable	Analysevariable
Zufrieden mit Lohn (Dummy)	a)-Variable	Analysevariable
Zeit in (geförderter) Beschäftigung bis Interviewzeitpunkt	a)-Variable	a)-Variable
Beschäftigung zum Interviewzeitpunkt beendet?	a)-Variable	Analysevariable
Nationalität deutsch (Dummy)	b)-Variable	b)-Variable
Berufserfahrung in Jahren	a)-Variable	a)-Variable
Nettohaushaltseinkommen	a)-Variable	Analysevariable

a)-Variable: Variable hängt potentiell mit der ausfallbelasteten Variable zusammen;
b)-Variable: Variable hängt potentiell mit dem Ausfall zusammen;
Analysevariable: Variable bereits im Analysemodell enthalten (vgl. Schafer 1999a: 143).

Anmerkung: Um das Vorgehen übersichtlich und verständlich zu halten, wird in der Untersuchung ein einziges Imputationsmodell für alle Szenarien verwendet. Dieses Modell wurde ursprünglich lediglich für die Imputation der Variable „Arbeitslosigkeitsdauer" aufgestellt (Szenario 2). Grundsätzlich müssten mit jeder neuen ausfallbelasteten Variable aber weitere a)-Variablen hinzugefügt werden, um die MAR-Annahme zu stützen. Das Imputationsmodell scheint sich aber – so zeigen die positiven Ergebnisse – auch ohne Erweiterung der a)-Variablen für die Szenarien 3 und 4 zu eignen. Für Szenario 1 gibt es hingegen definitionsgemäß keine a)-Variablen, da diese bereits im Analysemodell enthalten sein sollten.