Melanie Böwing-Schmalenbrock | Anne Jurczok

Multiple Imputation in der Praxis

Ein sozialwissenschaftliches Anwendungsbeispiel

Online veröffentlicht auf dem
Publikationsserver der Universität Potsdam:
URL http://opus.kobv.de/ubp/volltexte/2012/5811/
URN urn:nbn:de:kobv:517-opus-58111
http://nbn-resolving.de/urn:nbn:de:kobv:517-opus-58111

MULTIPLE IMPUTATION IN DER PRAXIS

EIN SOZIALWISSENSCHAFTLICHES ANWENDUNGSBEISPIEL

Autor: Melanie Böwing-Schmalenbrock; Anne Jurczok

Zusammenfassung

Multiple Imputation hat sich in den letzten Jahren als adäquate Methode zum Umgang mit fehlenden Werten erwiesen und etabliert. Das gilt zumindest für die Theorie, denn im Angesicht mangelnder anwendungsbezogener Erläuterungen und Einführungen verzichten in der Praxis viele Sozialwissenschaftler auf diese notwendige Datenaufbereitung. Trotz (oder vielleicht auch wegen) der stetig fortschreitenden Weiterentwicklung der Programme und Optionen zur Umsetzung Multipler Imputationen, sieht sich der Anwender mit zahlreichen Herausforderungen konfrontiert, für die er mitunter nur schwer Lösungsansätze findet. Die Schwierigkeiten reichen von der Analyse und Aufbereitung der Zielvariablen, über die Software-Entscheidung, die Auswahl der Prädiktoren bis hin zur Modell-Formulierung und Ergebnis-Evaluation. In diesem Beitrag wird die Funktionsweise und Anwendbarkeit Multipler Imputationen skizziert und es wird eine Herangehensweise entwickelt, die sich in der schrittweisen Umsetzung dieser Methode als nützlich erwiesen hat – auch für Einsteiger. Es werden konkrete potenzielle Schwierigkeiten angesprochen und mögliche Problemlösungen diskutiert; vor allem die jeweilige Beschaffenheit der fehlenden Werte steht hierbei im Vordergrund. Der Imputationsprozess und alle mit ihm verbundenen Arbeitsschritte werden anhand eines Anwendungsbeispiels – der multiplen Imputation des Gesamtvermögens reicher Haushalte – exemplarisch illustriert.

MULTIPLE IMPUTATION IN PRACTICE

A SOCIO-SCIENTIFIC EXAMPLE OF USE

Abstract

Multiple imputation established itself and proved adequate as method of handling missing observations – at least in theory. Annotations and explanations on how to apply multiple imputation in practice are scare and this seems to discourage many social scientists to conduct this step of necessary data preparation. Despite (or maybe because of) the continuous and progressive development of programs and features to conduct multiple imputation the user is confronted with numerous challenges for which solutions are sometimes hard to find. The difficulties range from the analysis and preparation of the target variable to deciding in favor of a software package, selecting predictors, formulating a suitable model and evaluating the results. This paper will outline the operation and practicability of multiple imputations and will develop a useful approach, which has proven adequate in handling missing values step by step – even for beginners. It will discuss potential difficulties and gives specific solutions; especially the particular quality of missing data is paramount. The process of imputation with all its necessary steps will be illustrated by the multiple imputation of the total assets of wealthy households.

Inhaltsverzeichnis

1. Einführung

1.1. Problembeschreibung

Es ist ein altbekanntes und unausweichliches Problem der empirischen Sozialwissenschaften: In jede Erhebung schleichen sich Messfehler ein. Sie entstehen entweder durch unzureichende oder ungenaue Angaben der Befragten, durch unsystematische Fragebogen- und Filterführung oder auch durch technische Fehler. Das Resultat solcher Messfehler sind zum einen unplausible und inkorrekte Werte, zum anderen unvollständige Angaben, so genannte „Missings" bzw. fehlende Werte. Zwar können durch sorgfältige und vorsorgliche Planung und Durchführung der Datenerhebung viele Fehler umgangen bzw. reduziert werden, sie lassen sich aber nicht vollständig vermeiden. Entscheidend für jede Analyse ist es, sich für sie zu sensibilisieren, sie zu identifizieren und einen Weg zu finden, ihren Einfluss auf die Messergebnisse möglichst gering zu halten. Denn werden inkonsistente oder fehlende Werte für grundlegende Variablen vernachlässigt und nicht korrigiert, schöpft der Datenbenutzer die Informationen der Daten einerseits nicht ausreichend aus und andererseits büßen die Daten und deren Analysen mitunter erheblich an Aussagekraft ein.

Zwei wesentliche potenzielle Gefahren können durch fehlende Werte ausgelöst werden bzw. dadurch, dass diese bei Analysen ignoriert werden (z.B. im Fall von „complete case analyses"): zum einen führt die verkleinerte Stichprobengröße zu einem Präzisionsverlust der Analysen. Zum anderen liegt zumeist ein Selektionsprozess der Forscher zugrunde (vor allem wenn die fehlenden Antworten nicht zufällig auftreten), durch welchen wiederum Verzerrungen und systematische Fehler auftreten können (vgl. Royston/Carlin/White 2009). Bei klassischen Verfahren im Umgang mit fehlenden Werten, in denen entweder die entsprechenden Fälle aus den Analysen ausgeschlossen werden (listwise deletion bzw. pairwise deletion), oder die vorhandenen Beobachtungen nach der Verteilung in der Bevölkerung gewichtet werden, sind daher nach heutigem Kenntnisstand statistisch völlig unzureichend und führen mitunter zu starken Verzerrungen der Schätzwerte (Lüdtke et al. 2007; Allison 2001; Briggs et al. 2002).
Um diese Gefahren zu umgehen und gleichsam die Datenqualität zu erhöhen und alle Informationen auszuschöpfen, sind zwei zentrale Korrekturschritte vorzunehmen: Erstens müssen unplausible Werte *editiert*, d.h. durch korrekte Werte ersetzt werden. Dies ist dann möglich, wenn sich anhand der vorhandenen Daten eindeutig nachweisen lässt, dass ein bestimmter Wert nicht konsistent ist, und wenn sich darüber hinaus aus den übrigen Daten ein plausibler Wert herleiten lässt. Der zweite Korrekturschritt ist unlängst aufwändiger, denn in ihm werden fehlende Werte nicht direkt aus anderen Angaben abgelesen, sondern mit Hilfe entsprechender Annahmen und Verfahren geschätzt. Dabei handelt es sich um die *Imputation* der fehlenden Angaben. In den angewandten Wirtschafts- und Sozialwissenschaften hat sich mit dem Fortschreiten der technischen und statistischen Möglichkeiten eine Vielzahl verschiedener Imputationsmethoden entwickelt, von denen sich übereinstimmend insbesondere die Multiple Imputation als geeignet erwiesen hat.

Es fehlt allerdings an anwenderfreundlichen Erläuterungen zu den einzelnen Analyse- bzw. Arbeitsschritten im Verlauf einer Imputation. Es besteht ein dringender Bedarf an anwendungsorientierter Literatur für entsprechende Analysen (vgl. Kenward/Carpenter 2007: 215). So sieht sich die breite Mehrheit der potenziellen Anwender bei dem Versuch einer Umsetzung der Multiplen Imputation häufig mit schwer überwindbaren Problemen konfrontiert: Erstens bildet die auffindbare Literatur überwiegend nur die statistischen Hintergründe ab, kaum aber erfasst sie die Situation innerhalb der Datenanalyse. Zweitens sind verschiedene methodische Schwierigkeiten nicht in den Statistikprogrammen implementiert und entsprechende Lösungsansätze spärlich dokumentiert.

Dieser Beitrag verfolgt das Ziel, Sozialwissenschaftlern den Zugang zur konkreten Anwendung und Umsetzung einer Multiplen Imputation zu erleichtern, indem beispielhaft und anwendungsbezogen eine Annäherung an die Umsetzung einer Multiplen Imputation erörtert wird. Anhand eines Anwendungsbeispiels zur Imputation des Haushaltsgesamtvermögens in der Studie „Vermögen in Deutschland“ (*ViD*) mit Hilfe des benutzerdefinierten Befehls „ICE“ im Statistikprogramm STATA nähern wir uns an die methodischen Schritte an, skizzieren potenziell auftretende Schwierigkeiten und nennen mögliche Lösungswege. Wir erheben dabei keineswegs den Anspruch, dass unsere Darstellungen vollständig und unsere Vorschläge die einzigen oder gar besten sind. Sie sind lediglich als Dokumentation und potenzielle Hilfestellung gedacht. Multiple Imputation ist und bleibt ein aufwändiges und problembehaftetes Unterfangen, das noch am Anfang seiner Entwicklung steht. Nichtsdestoweniger ist sie die zurzeit bestgeeignete Methode zum Umgang mit fehlenden Angaben, und die nachfolgenden Ausführungen können hoffentlich dazu beitragen, sie auch in der sozialwissenschaftlichen Praxis stärker zu verbreiten.

1.2. Multiple Imputation

Allen Imputationsverfahren[1] ist gemein, dass die fehlenden Werte durch plausible Werte ersetzt werden. Welche Werte dabei eingesetzt werden, wird allerdings verschiedenartig gehandhabt. Die Auswahl eines Imputationsverfahrens sowie die Formulierung und Programmierung des entsprechenden Imputationsmodells sind von den konkreten Begebenheiten des Datensatzes und den Zielen der Anwender abhängig. Der Prozess der Imputation muss daher stets an die jeweilige Situation angepasst werden: „The method of imputation should depend on the context and available covariate data“ (Carlin et al. 2003: 226).

Eine angemessene Imputations-Methode sollte in der Lage sein, „to inject the correct degree of randomness into the imputations and to incorporate that uncertainty when computing standard errors and confidence intervals for parameters of interest.“ (Royston 2004: 228). Konventionelle Vorgehensweisen wie beispielsweise die Mittelwert-Imputation erfüllen diese Kriterien überwiegend nicht, sie verletzen zum einen häufig die verteilungsbasierte Zufälligkeit der Werte, zum anderen werden

[1] Eine ausführliche Auflistung von imputationsbasierten Verfahren bieten beispielsweise Briggs et al. 2002; Allison 2001; Lüdtke et al. 2007.

Standardfehler nicht oder nicht hinreichend berücksichtigt. Auf diese Weise können zwar alle Fälle verwendet werden, die wahre Verteilung wird indes mitunter stark verzerrt.

Das Verfahren der *Multiplen Imputation* erfüllt die Gütekriterien und ist – sofern die Daten dies ermöglichen – allen klassischen Imputations-Methoden vorzuziehen. Bei der Multiplen Imputation werden für die fehlenden Werte Schätzwerte eingesetzt, die durch die Verteilung verschiedener Prädiktoren vorhergesagt werden. Dies geschieht, indem alle vorliegenden relevanten Informationen des Datensatzes berücksichtigt und Zufallsfehler hinzugerechnet werden. Die Auswahl der verwendeten Prädiktoren richtet sich nach ihrer theoretischen und statistischen Relevanz zur Vorhersage des zu imputierenden Merkmals (van Buuren/Boshuizen/Knook1999; siehe Punkt 4.1). Dieser Schätzprozess wird mehrmals durchgeführt. Die einzelnen entstehenden Schätzwerte können schließlich für weitere Analysen herangezogen werden, bzw. ergibt sich durch sie der für anschließende Berechnungen bereitgestellte Wert. (Vgl. Rässler/Rubin/Zell 2007; Drechsler 2010)
Der unbestreitbare Vorteil der Multiplen Imputation ist einerseits der geringe Informationsverlust, da alle Variablen in das Modell mit einbezogen werden, die im Zusammenhang mit der zu imputierenden Variablen stehen oder die sich auf das Antwortverhalten der Befragten auswirken. Andererseits besticht diese Methode durch die Einbeziehung von Standardfehlern, die zudem durch die mehrfache (*m*-malige) Wiederholung des Schätzprozesses zufällig und realistisch berechnet werden. Im Ergebnis erhält man jeweils *m* Werte aller imputierten (und nun vollständigen) Merkmale, die entweder allesamt in weitere Analysen einfließen oder zu einem einzelnen Wert kombiniert werden können.

Statistisch hat sich die Methode der Multiplen Imputation bereits als äußerst zuverlässiges Verfahren bewiesen: Spieß und Göbel beispielsweise finden für Einkommensangaben heraus, dass „the results based on multiply imputed data sets are more reliable than those based on the complete case analysis“ (2005: 63). Ähnlich stellen Marchenko und Reiter (2009: 388) fest, dass multivariate Analysen mit multipel imputierten Daten Vorteile gegenüber Analysen mit nicht angepassten Daten haben: „We propose improvements to existing degrees of freedom used for significance testing of multivariate hypotheses in small samples when missing data are handled using multiple imputation.“ Vor allem bei einer stärkeren Selektivität des Datenausfalls „lieferte ausschließlich die Multiple Imputation gute Ergebnisse“ (Krug 2010: 28). Ebenso überzeugt einzig dieses Verfahren beim Umgang mit fehlenden Werten von Variablen zu Finanzfragen, wie dies bereits seit mehreren Jahren von den Analysten des amerikanischen „Survey of Consumer Finances“ vorgelebt wird (Kennickell 1998). In den vergangenen Jahren wurden Multiple Imputationsverfahren zu „an important and influential approach in the statistical analysis of imcomplete data“ (Kenward/Carpenter 2007: 199).

Multiple Imputation ist schlichtweg aktuell das am besten geeignete und auch von Kritikern bevorzugte statistische Verfahren zur Handhabung fehlender Werte (Allison 2001; Schafer et al. 2002; Rubin 2004; Horton et al. 2003, 2007; Lüdtke et al. 2007; Fessler et al. 2009).

2. Das Anwendungsbeispiel

Zielsetzung und Analyse der fehlenden Werte

Zu Beginn einer jeden Imputation muss der Forscher sich bewusst machen, welche Variablen imputiert werden sollen, wie viele fehlende Werte es bei dieser Variable gibt und warum die Werte fehlen. Eine genaue Kenntnis der zu imputierenden Variablen – auch Zielvariablen – und des Datensatzes gelten dabei als Voraussetzung. Im vorliegenden Fall besteht das Ziel der Imputation darin, eine Variable zu erstellen, in der vollständige und zuverlässige Informationen über die Höhe des Gesamtvermögens für die Studie „Vermögen in Deutschland" (*ViD)* vorliegen. Anders als häufig üblich, sollen nicht mehrere imputierte Datensätze erstellt werden, die an einem konkreten Analysemodell ausgerichtet sind, sondern der Imputationsprozess und das zugehörige Modell sollen einen allgemeingültigen Punktschätzer für das Haushaltsgesamtvermögen in *ViD* ausweisen, der für verschiedenste Fragestellungen und Analysen verwendbar ist.

ViD ist eine quantitative Untersuchung, in der gezielt reiche Haushalte zu verschiedenen Lebensbereichen befragt wurden.[2] Das Gesamtvermögen der Haushalte wurde in *ViD* zweifach abgefragt, zum einen als offene Angabe, bei der die Befragten den von ihnen geschätzten Wert direkt angeben. Diese Variable ist es auch, deren Imputation nachfolgend dokumentiert wird. Zum anderen verbirgt sich die Höhe des Gesamtvermögens hinter der Summe der Geldanlagen auf der einen und der sonstigen Vermögensbestände auf der anderen Seite. Diese beiden Variablen wurden kategorial abgefragt (Fragebogenausschnitte finden sich im Anhang). Durch den Abgleich mit den Kontrollfragen können die Angaben auf interne Konsistenz geprüft und gegebenenfalls editiert werden (siehe 3.1).

Fragen zur Einkommens- oder Vermögenssituation sind generell heikel und lassen sich deshalb schwer verlässlich erheben (vgl. Riphahn/Serfling 2002; Frick/Grabka 2009). Das hat mehrere Ursachen: zum einen kennen viele Menschen diese Werte schlichtweg nicht, weil sie zum Befragungszeitraum keinen genauen Überblick über ihre Vermögenswerte haben. Vor allem Immobilienvermögen, oder der Gegenwert von Sach- oder Betriebsvermögen können zudem oft nur grob geschätzt werden. Da in Haushalten mit hohen Vermögen häufig viele verschiedene Vermögensarten vorliegen, potenzieren sich diese Unsicherheiten hier sogar noch. Zum anderen ist das Sprechen über Geld, in Deutschland ein weitgehend tabuisiertes Thema. „Hohe Einkommensbezieher lieben das Diskrete, vor allem, wenn es um die Offenbarung ihrer Einkünfte geht" (Huster 2009: 45).

Trotz verschiedener präventiver Maßnahmen innerhalb der Fragebogengestaltung, Erhebungsart und des Gesamtdesigns der Studie *ViD* konnten fehlende Werte und Inkonsistenzen in den Vermögensangaben nicht vollständig verhindert werden. Für 135 Fälle liegt kein gültiger Wert für die Höhe des Haushaltsgesamtvermögens vor. Diese Anzahl entspricht 28,6 Prozent an allen befragten Haushalten (vgl. Tabelle 1). Ausfallquoten in dieser Größenordnung sind für Vermögensangaben durchaus üblich. Im Sozio-Ökonomischen Panel beispielsweise lagen die Ausfallquoten der offen abgefragten Vermö-

[2] Verantwortlich für die Durchführung der Studie *ViD* sind Prof. Dr. Wolfgang Lauterbach und Melanie Böwing-Schmalenbrock (Universität Potsdam). Die Studie wurde in Kooperation mit Prof. Dr. Thomas Druyen (SFU Wien) und Prof. Dr. Matthias Grundmann (WWU Münster) und in Zusammenarbeit mit TNS Infratest Sozialforschung durchgeführt. Weitere Informationen zur Studie sind nachzulesen bei Lauterbach/Kramer/Ströing 2011 sowie Kortmann 2011.

genswerte je nach Vermögensbestandteil zwischen 25 und 50 Prozent (vgl. Schäfer/Schupp 2006; Frick/Grabka 2009). Im Imputations-Prozess erhöht sich zwar mit dem Anteil an fehlenden Werten auch die Unsicherheit der Schätzverfahren, bis zu Ausfallquoten von 50 Prozent ist eine Imputation jedoch nicht nur machbar, sondern auch empfehlenswert, da durch sie die Analysen auch bei derart vielen fehlenden Werten zuverlässiger werden als ohne Imputation (vgl. Royston 2004: 240).

Tabelle 1: Übersicht über vorhandene Angaben zu Vermögensvariablen

	Fälle	Umgang mit Haushalts-gesamtvermögen
Vollständig vorhandene Information zu allen Vermögensfragen	312 (66,1%)	Konsistenztests, ggf. Editierung; keine Imputation
Das offen abgefragte Gesamtvermögen liegt vor, die übrigen Vermögensangaben sind unvollständig	25 (5,3%)	Bleibt unverändert
Fehlende Fälle beim offen erfragten Haushaltsgesamtvermögen	135 (28,6%)	Imputation
Darunter: …vollständig vorliegende Angaben zu kategorial abgefragten Vermögensbeständen	*75 (15,9%)*	
…unvollständige Angaben zu kategorial abgefragten Vermögensbeständen	*60 (12,7%)*	
Gesamt	**472 (100%)**	

3. Methodische Vorbereitungen und Voraussetzungen

3.1. Editing

Da im Verlauf des Imputations-Prozesses die existierenden Werte aller verwendeten Merkmale als Schätz-Grundlage dienen, ist es umso bedeutsamer, sich der Plausibilität und dem Wahrheitsgehalt der Daten möglichst sorgsam zu vergewissern. Bevor mit den statistischen Analyseschritten begonnen werden kann, sollte daher sichergestellt werden, dass die vorhandenen beobachteten Werte in sich stimmig und untereinander konsistent sind. Denn alle weiteren Analyseschritte basieren auf der Verteilung der vorhandenen Werte, sodass Fehler in der Datensubstanz unter Umständen gravierende Auswirkungen hätten. Werden im Zuge dieser Überprüfungen bestehende Werte angepasst oder können fehlende Angaben bereits durch logische Schlussfolgerungen ersetzt werden, handelt es sich um eine Editierung der Daten.

Für die *Plausibilitätsprüfung* können die Ausreißer der relevanten Variablen betrachtet und die Verteilung mit vergleichbaren amtlichen Daten verglichen werden oder die interne Konsistenz der Variable überprüft werden. Da besonders hohe und besonders niedrige Werte den Schätzprozess sensibel beeinflussen, muss sichergestellt werden, dass es sich bei ihnen um keine Messfehler handelt. Die Ausreißer-Analyse stellt bereits eine zielgerichtete Form eines Tests auf interne Konsistenz der Daten dar. Dabei gibt das Zusammenspiel verschiedener Variablen Hinweise darauf, ob wahrheitsgemäße bzw. plausible Werte vorliegen. Werden bestimmte Informationen in einer Erhebung auf

unterschiedliche Weise erfragt, können die einzelnen Items aufeinander abgestimmt werden. Sogenannte Kontrollfragen werden eben zu diesem Zweck in vielen komplexen Befragungen integriert. Jeder Datensatz sollte gründlich auf entsprechende Vergleichbarkeiten hin untersucht werden. Für die spätere Modellformulierung ist dieser Schritt ohnehin notwendig.
In einigen Fällen können fehlende Werte bereits eingesetzt werden, ohne eine statistische Schätzung vorzunehmen. Solche *logischen Imputationen* sind vor allem dann sinnvoll, wenn Angaben systembedingt fehlen, etwa aufgrund der Filterführung. Bevor mit der Imputation begonnen wird, sollte daher differenziert werden, welche Missings systemisch zulässig sind und welche nicht. Für das hier beschriebene Imputations-Verfahren ist es notwendig, die beobachteten Werte aller relevanten Variablen zu überprüfen. Beispielhaft wird dieser Vorgang an der zu imputierenden Angabe, dem offen abgefragten Gesamtvermögen der Haushalte ausgeführt[3]:

Beispiel

Um zu überprüfen, ob es sich bei den beobachteten Werten um realistische handelt, können beispielsweise Verteilungen aus amtlichen Daten herangezogen und mit den eigenen Daten verglichen werden. Da aufgrund der hohen Vermögen in der Stichprobe keine repräsentativen amtlichen Vergleichsdaten vorliegen, beschränkt sich die Plausibilitätsprüfung im vorliegenden Fall überwiegend auf interne Tests. Eine Zusammenfassung der Editierung ist *Tabelle 2* zu entnehmen.

Tabelle 2: Fallkonstellationen nach Konsistenzprüfung

Situation	Fälle	Vorgehensweise bei Editing
Alle drei Variablen vorhanden und konsistent	183	Generierte Variable = Ursprungsvariable
Ursprungsvariable vorhanden und >= 200.000, keine Konsistenzprüfung möglich, da v3.5 und/oder v3.9 fehlen	25	Generierte Variable = Ursprungsvariable
Ursprungsvariable ist geringer als Minimalsumme aus v3.5 und v3.9	106	Generierte Variable = Summe der mittleren Werte aus v3.5 und v3.9
Ursprungsvariable ist höher als Maximalsumme aus v3.5 und v3.9	21	Generierte Variable = Ursprungsvariable
Keine Angabe bei Ursprungsvariable	**135**	**Generierte Variable = missing** (wird imputiert)
Trotz Generierung liegt generierte Variable unter 200.000	**2**	**Generierte Variable = missing** (wird imputiert)
Gesamt	*472*	

Legende: Ursprungsvariable = offen abgefragtes Haushaltsgesamtvermögen (v3.10); generierte Variable = Ursprungsvariable nach Editing (vermgen); v3.5 = Höhe des Geldvermögens, kategorial; v3.9 = Höhe sonstiger Vermögensbestände, kategorial. Die Frageformulierungen und Kategorien können im Anhang eingesehen werden.

Die *Analyse der Ausreißer* beim offen erfragten Gesamtvermögen der Haushalte ergab, dass einige Angaben unplausibel sind. Werte unter 200.000€ sind bereits aufgrund des Selektionskriteriums der

[3] Für eine ausführliche Darstellung eines Editing-Verfahrens bei Einkommens- und Haushaltsangaben siehe Frick/Grabka 2007.

Stichprobenauswahl nicht möglich und werden daher gelöscht, sofern sie nicht editiert werden können. Ausnahmen bilden hierbei einzelne Werte, für die mit sehr hoher Wahrscheinlichkeit angenommen werden kann, dass bei der Eingabe lediglich eine oder mehrere Nullen vergessen wurden.[4] Für die einzelnen Ausreißer nach oben konnten keine Unplausibilitäten festgestellt werden, zumal es sich jeweils um Haushalte mit Betriebsvermögen handelt.

Eine andere Möglichkeit der Plausibilitätsprüfung der Daten ist es, die Ausreißer mit den übrigen Angaben des jeweiligen Falles abzugleichen, um so ein Eindruck zu gewinnen, ob die Werte plausibel sind. In der Vermögensstudie wurde die *interne Konsistenz* der Variable vor allem durch einen Abgleich zwischen der offen abgefragten metrischen Vermögensvariable und den beiden kategorial erfragten Vermögensvariablen geprüft werden. Denn aus der Summe letzterer erschließt sich die Höhe des Gesamtvermögens. Wird diesbezüglich eine Inkonsistenz festgestellt, werden die zugehörigen Werte des Haushaltsgesamtvermögens – sofern möglich – editiert. Diese Editierung sollte in Syntaxbefehlen formuliert werden, manuelle Tests und Eingaben wären vor allem für größere Datensätze nicht realisierbar.

Tabelle 3: Eckwerte der Verteilung des Haushaltsgesamtvermögens vor und nach dem Editing

		v3.10 (Ursprungsvariable)	**vermgen (Generierte Variable)**
N	Gültig	337	335
	Fehlend	135	137
Mittelwert		**1.965.104**	**2.289.866**
Standardabweichung		3.992.626	4.036.966
Minimum		2	200.000
Maximum		50.000.000	50.000.000
Perzentile	25	500.000	750.000
	50 (Median)	**900.000**	**1.200.000**
	75	2.000.000	2.250.000

Es gilt zu beachten, dass während der Konsistenz- und Plausibilitätsprüfung noch keine fehlenden Werte imputiert werden, dies geschieht erst durch statistische Schätzverfahren im nächsten Schritt (siehe 3.2). Durch die Editierung der Inkonsistenzen verschiebt sich die durchschnittliche Vermögenshöhe im Anwendungsbeipsiel automatisch nach oben, da zumeist die höhere Angabe bevorzugt wurde, sofern die Angaben nicht konsistent waren (vgl. *Tabelle 3*). Allein das Ersetzen der einziffrigen Werte durch die entsprechenden Millionenbeträge (siehe Fußnote 4) hat einen starken Effekt auf die durchschnittliche Vermögenshöhe, sodass diese höheren Verteilungswerte nicht verwunderlich sind.

[4] Elf Haushalte in *ViD* nennen ein Gesamtvermögen von weniger als 200.000 Euro. Bei einziffrigen Nennungen kann aufgrund der übrigen Vermögensangaben mit sehr großer Wahrscheinlichkeit angenommen werden, dass es sich um Millionenbeträge handelt (zwei Fälle betroffen). Die übrigen Ausreißer nach unten können überwiegend durch das standardisierte Editing (s.u.) ersetzt werden, da aufgrund der vorliegenden Werte für das Kapitalvermögen und die sonstigen Vermögensbestände bekannt ist, dass die Vermögen hier jeweils höher liegen. Sofern auch nach dem Editing kein Wert über 200.000 vorliegt, wird ein Missing gesetzt; das ist in zwei Fällen der Fall.

Es kann insgesamt davon ausgegangen werden, dass diese Korrektur nicht zu einer Überschätzung der tatsächlichen Vermögensbestände führt, sondern dass vielmehr ohne die Korrektur der beobachteten Vermögensangaben tendenziell die tatsächlichen Bestände unterschätzt worden wären.

3.2. Voraussetzungen für die Durchführung einer Multiplen Imputation

Normalverteilung der Zielvariablen

Der Imputations-Prozess ist konzipiert für normalverteilte Variablen (Lee/Carlin 2010). Die Normalverteilungs-Bedingung wird allerdings nicht von allen Programmen und auch längst nicht mehr von allen Forschern derart streng behandelt (Schafer 1997, siehe auch Lee/Carlin 2010). Es ist von verschiedenen Kniffen zu lesen, die gewährleisten, dass zumindest eine gelockerte Normalverteilungsannahme nicht verletzt wird. Ein besonders effektiver Kniff empfiehlt sich für metrische Variablen: Ihre Verteilung kann durch Logarithmierung korrigiert und damit der Einfluss der eigentlichen Schiefe reduziert werden: „Log transformation makes the data more normal" (Royston 2007: 452)[5]. Die „log-skew0" Transformation versucht im Gegensatz zum normalen Logarithmieren eine 0-schiefe Verteilung herzustellen und produzierte in einer Simulation geringere Verzerrungen bei den imputierten Werten, als die einfache Log-Transformation.[6] Zusätzlich kann in dem statistischen Modell, z.B. im Befehl ICE, die „bootstrap" Methode angeführt werden, durch welche die Normalverteilungsannahme nochmals gelockert wird (vgl. Royston 2004: 232).

Auch das Gesamtvermögen aus dem Anwendungsbeispiel ist nicht normal verteilt, vielmehr liegt eine rechtsschiefe Verteilung vor. Aus diesem Grund wird eine neue Imputation-Variable erstellt (lnvermgen), welche die logarithmierten Werte der ursprünglichen Vermögensvariablen (vermgen) enthält. Im Anschluss an die Imputation können die imputierten Variablen rückgängig transformiert, also „delogarithmiert" werden, um den ursprünglichen Wertebereich wieder herzustellen.[7]

Bestimmung des Missing-Mechanismus

Der Umgang mit fehlenden Werten und die Auswahl eines Imputationsverfahrens hängen davon ab, wie hoch die Wahrscheinlichkeit ist, ob die zu schätzenden Werte zufällig oder nicht zufällig fehlen. Dazu werden Informationen über die zu imputierende Variable, die abhängigen Variablen und unberücksichtigte Variablen, also solche die das Antwortverhalten des Befragten beeinflussen, benötigt (vgl. McKnight et al. 2007: 42 ff., Göthlich 2007: 121 oder Krug 2010: 30ff). Es wird zwischen drei Missing-Mechanismen unterschieden: Missing Completely at Random (MCAR), Missing at Random (MAR) und Missing not at Random (MNAR). Je nachdem welcher Missing-Mechanismus vorliegt, kann es zu unterschiedlichem Grad von Verzerrungen im Imputationsprozess und zu Fehlinterpretation der imputierten Daten kommen (Spieß 2010: 118).Wenn die Werte völlig zufällig fehlen, also das Fehlen nicht abhängig ist von den beobachteten oder anderen nicht beobachteten Werten und dem Muster der fehlenden Werte (missing pattern), dann werden die fehlenden Werte als *Missing Completely at*

[5] Weitere Hinweise zur Nützlichkeit von logarithmierten Variablen bei Allison (2002: 39) oder STATA Press (2009: 10).
[6] Siehe Lee/Carlin 2010 für die ausführliche Erklärung des Verfahrens.
[7] Beispiel für ICE (STATA): Einfache Log-Transformation: generate lnvermgen = ln(vermgen) oder bei der „log-skew0"-Transformation: lnskew0 lnvermgen=vermgen. Für die Delogarithmierung: generate vermgen_1 = exp(lnvermgen).

Random (MCAR) bezeichnet. Davon zu unterscheiden ist der Mechanismus *Missing at Random (MAR),* bei dem das Muster der fehlenden Werte mit den beobachteten Werten des gleichen Merkmals zusammenhängt. Das bedeutet also, dass das Antwortverhalten für eine bestimmte Variable von der tatsächlichen Ausprägung derselben Variable beeinflusst wird. Sobald das Muster der fehlenden und beobachteten Werte mit den beobachteten UND unbeobachteten Werten zusammenhängt, liegt der *Missing Not at Random (MNAR) Mechanismus* vor (Göthlich 2007: 121; McKnight et al. 2007: 49). Der MNAR Mechanismus ist anders als der MAR und MCAR nicht ignorierbar ohne einen Präzisionsverlust bei der Imputation zu gefährden.: "When the mechanism is nonignorable and the amount of missing data is not trivial, any alternative is questionable that fails to include the variables or the mechanism that account for the missing data" (McKnight et al. 2007: 128).
Die Unterscheidung zwischen MNAR und MAR ist schwierig, da der einzige Merkmalsunterschied die Abhängigkeit des Missing-Mechanismus bei MNAR von den Ausprägungen der nicht-beobachtbaren, fehlenden Werte ist. Da man diese Merkmalsausprägungen nicht kennt, ist es schwer ihre Abhängigkeit nachzuweisen. McKnight et al. (2007: 95) schreiben, dass "there is no diagnostic procedure, numeric or graphic, that validly differentiates between MAR and MNAR. Instead, we must rely on logic and a sound understanding of the study design and domain". Im Idealfall können Angaben derselben Befragten aus früheren oder späteren Erhebungen herangezogen werden, ansonsten beispielsweise Informationen über die Verteilung des entsprechenden Merkmals aus anderen Studien oder amtlichen Daten. Liegen entsprechende Daten nicht vor dann muss der Forscher sich auf seine eigene Logik stützen (vgl. McKnight et al. 2007: 98; Schafer 1997: 22). Liegt der Verdacht vor, dass die zu imputierenden Werte nicht zufällig sind (MNAR), sollten möglichst viele Prädiktoren, die in Abhängigkeit mit der Zielvariable stehen oder Einfluss auf die Responsewahrscheinlichkeit haben, in das Imputationsmodell eingeschlossen werden (Spieß 2010: 126; Kenward/Carpenter 2007: 205) (siehe 4.1). Trotzdem, selbst wenn Imputationen bei einem Missing Data Mechanismus von MNAR durchgeführt werden, müssen die Ergebnisse nicht zwangsläufig verzerrt sein, insbesondere wenn es sich um binäre Merkmalstypen handelt (Hohl 2008: 128). Entscheidend ist lediglich, dass alle verfügbaren Informationen berücksichtigt werden, die über die Verteilung der Zielvariablen und das Antwortverhalten vorliegen, und der Einfluss des Missing-Mechanismus möglichst gering gehalten wird. Alle Möglichkeiten sollten analysiert und alle entsprechenden Variablen in das Imputationsmodell aufgenommen werden. In der Praxis bedeutet dies vor allem, den Zusammenhang zwischen der Ziel- und aller in Betracht kommender abhängigen Variablen zu untersuchen sowie die Verteilung der beobachteten Werte der Zielvariablen wenn möglich mit der wahren Verteilung abzugleichen.

3.3. Auswahl der Software

Multiple Imputation kann mittlerweile mit verschiedenen Statistik Programmen durchgeführt werden, die teilweise wiederum unterschiedliche Software Pakete anbieten, also verschiedene Herangehensweisen. [8] Es ist nicht unbedingt nötig, den gesamten Imputationsprozess mit einem einzigen

[8] Für eine Auflistung unterschiedlicher Imputations-Routinen und Programme siehe Horton et al. 2007, Lüdtke et al. 2007, Hohl 2008: 48ff., Spieß 2010.

Software Paket zu durchlaufen. Allerdings bietet es sich an, sich zumindest auf ein Statistik Programm festzulegen. Die nachfolgenden Ausführungen beschränken sich auf das Progamm STATA. Die generellen Handhabungen können auch auf andere Programme, insbesondere SPSS übertragen werden, im Detail jedoch richten sie sich eng nach den methodischen Implementationen innerhalb von STATA. Innerhalb von STATA gibt es im Wesentlichen zwei Möglichkeiten Multiple Imputationen durchzuführen: Einerseits kann mithilfe des benutzergeschriebenen Software Pakets ICE (Royston 2004, 2005a, 2005b, 2007, et al. 2009) zum anderen mit den STATA eigenen mi-Befehlen (STATA 11) eine Multiple Imputation durchgeführt werden. Während ICE auf einem Imputationsalgorithmus basiert, wonach der Imputationsprozess aus vielen einzelnen univariaten Imputationen (deshalb auch „chained equation" genannt) besteht, die in einem Modell zusammengefasst werden (vgl. van Buuren/Boshuizen/Knook 1999; van Buuren 2007), stützen sich die mi-Befehle von STATA 11 auf einem Algorithmus, bei dem unter der Annahme einer multivariaten Normalverteilung der Daten ein Modell für alle Variablen formuliert wird (vgl. Schafer 1997). Unterschiede gibt es außerdem hinsichtlich der Berücksichtigung unterschiedlicher Merkmalstypen von Variablen und dem Umgang mit normalverteilten Daten. Im Folgenden werden beide Herangehensweisen erläutert und es wird ausgeführt, wie sie derart verknüpft werden können, dass jeweils für jeden Arbeitsschritt die passende Methode verwendet wird.

Trotzdem STATAs Mi-Befehle und ICE auf unterschiedlichen Routinen basieren, können sie in der Praxis komplementär verwendet werden. Die Mi-Kommandos sind theoretisch fundierter und ausführlicher beschrieben und können besonders in dem Analyse- und Poolingschritt, also beim Zusammenspielen der unterschiedlichen imputierten Datensätze, hilfreich sein. Auch die Analyse des Missing-Mechanismus und des Missing-Musters ist besonders einfach in STATAs mi-Kommandos durchführbar. Folglich bedeutet das, dass es sinnvoll ist, die Daten für die anfängliche Analyse in die mi-Kommandos zu konvertieren („mi set"). Anschließend können die möglichen Analysen der Missing-Mechanismen[9] durchgeführt werden. Das eigentliche Imputationsmodell, also das Herzstück der Imputation, sollte nach unserer Empfehlung mit Hilfe der Befehle von ICE formuliert werden. Hierfür ist es notwendig die Daten von mi nach ICE zu importieren („mi export ICE"). Der abschließende Analyseschritt, bei dem die Güte des Imputationsmodells überprüft wird, ist ebenfalls einfacher mit den mi-Kommandos[10] umzusetzen, weil komplexe Routinen vorhanden sind und die einzelnen Befehle innerhalb eines Handbuches ausführlich beschrieben werden. Zur Umsetzung dieser Schritte äußert sich das folgende Kapitel.

4. Imputations-Prozess

Für die Analysen wird es sicherlich notwendig sein, zuvor einige Features über die STATA Plattform herunterzuladen oder zumindest zu aktualisieren. Da nicht alle Optionen in dem Ursprungsprogramm automatisch enthalten sind, etwa weil sie wie im Beispiel von ICE von Nutzern zur Verfügung gestellt

[9] z.B. „mi misstable sum", „mi misstable nested", „mi misstable patterns", „mi misstable tree"
[10] z.B. „mi xeq", „mi estimate", „vartable:regress"

und stetig weiterentwickelt werden, muss man sie zunächst über das Kommando „findit“ suchen und sie schließlich herunterladen[11]. Das gilt aktuell für die Optionen ICE (für den Imputationsschritt) und mim (für den Analyse- und Poolingschritt), kann aber stetig um weitere aktuelle Kommandos erweitert werden.
Als Herangehensweise an die Formulierung des statistischen Modells und damit an das Herz der Imputation hat es sich in unserem Fall als äußerst sinnvoll erwiesen, das STATA Handbuch über Multiple Imputation zur Hand zu nehmen (STATA Multiple-Imputation Reference Manual Release 11) und die zugehörigen Kapitel in der vorgeschlagenen Reihenfolge durchzuarbeiten. Für die Nutzung von ICE empfiehlt es sich die Artikel von Royston et al. (2004, 2005a, 2005b, 2007, 2009) aus dem STATA Journal zu lesen. Nachfolgend werden einige sowohl allgemeine Analyseschritte als auch spezielle Problemstellungen dargelegt und es werden jeweils Vorschläge für den Umgang mit ihnen gemacht.

4.1. Auswahl der Prädiktor-Variablen

Der vermutlich entscheidendste Schritt bis zum Imputationsmodell ist die Auswahl der Variablen, anhand derer die imputierten Werte geschätzt werden, der Prädiktoren.
In das Imputationsmodell müssen alle Variablen aufgenommen werden, die mit der zu imputierenden Variable im Zusammenhang stehen oder auf das Antwortverhalten der Befragten Einfluss nehmen. Auch so genannte Hilfsvariablen können die Effizienz des Imputationsmodelles erhöhen, „the inclusion of these variables is at worst neutral, and at best extremely beneficial” (Collins/Schafer/Kam 2001: 348). „Es ist deshalb ratsam, neben den eigentlich interessierenden Variablen möglichst viele zusätzliche Variablen, die im Zusammenhang mit dem Ausfallprozess stehen, in den Datensatz aufzunehmen“ (Lüdtke et al. 2007: 105). Auch Variablen, die die generelle Motivation zur Teilnahme an Untersuchungen abfragen, sind in diesem Zusammenhang sinnvolle Prädiktoren, da diese Hinweise auf das generelle Antwortverhalten aufzeigen und somit helfen können die MNAR Annahme zurückzuweisen (Spieß 2010: 126). Für den Fall, dass später mit den verschiedenen im Laufe des Imputationsprozesses geschätzten Werten ein konkretes Modell gerechnet werden soll und die Imputation an diese konkrete Analyse angepasst wird, muss zudem darauf geachtet werden, alle für die geplante Analyse relevanten Variablen auch bereits bei der Imputation zu verwenden. Insgesamt ist es zweckmäßig, ein möglichst großes Imputationsmodell zu wählen, denn so können die einzelnen Annahmen entspannter betrachtet werden (siehe Royston 2007: 461; STATA Press 2009). Dennoch sollte die Auswahl nicht willkürlich ablaufen.

Einen geeigneten Vorschlag zur schrittweisen Vorgehensweise bei der Prädiktoren-Auswahl machen beispielsweise van Buuren/Boshuizen/Knook (1999):

1. Zunächst wird ein theoretisches Modell aufgespannt, in das aus den vorhandenen Variablen des Datensatzes all jene ausgewählt werden, die einen theoretischen Zusammenhang mit der zu imputierenden Variable aufweisen („U“) oder Einfluss darauf haben, ob ein fehlender Wert vorliegt oder nicht („V“).

[11] Kommandos für ice: st0067_4/ice_; mim:st0139_1; mvpattern: dm91

2. Von dieser Variablenliste verbleiben weiterhin diejenigen im Modell, für die sich auch statistisch dieser Zusammenhang bestätigt (signifikantes Chi²).
3. „Usable Cases“: Wenn in weniger als 50 Prozent der Fälle, in denen die Zielvariable einen fehlenden Wert hat, gültige Werte der Prädiktor-Variablen vorhanden sind, ist diese aus dem Modell zu löschen.

Da aber bei der Multiplen Imputation alle verwendeten Variablen zunächst imputiert werden, bevor sie zur Imputation des Vermögens herangezogen werden (sofern nicht im Modell anders formuliert), sollte für jede einzelne verwendete Variable mit fehlenden Werten ein Imputationsmodell innerhalb des Gesamtmodells formuliert werden. Die Auswahl der Prädiktoren erfolgt dabei analog zu dem soeben für das Vermögen beschriebenen Auswahlprozess, es werden also jeweils nur Prädiktoren ausgewählt, die mit der zu imputierenden Variablen einen signifikanten Zusammenhang aufweisen, der theoretisch begründet werden kann. Im eigentlichen Modell können dann mehrere „Untermodelle“ formuliert werden, aus denen die Wechselwirkungen der Prädiktoren untereinander jeweils hervorgehen.

Beispiel

Im Anwendungsbeispiel werden zunächst alle Variablen getestet, für die ein entsprechender Einfluss auf das Haushaltsgesamtvermögen oder das Antwortverhalten vermutet wird. Nach diesem Schema verbleiben ausschließlich Prädiktoren im Modell, die entweder mit der Zielvariablen direkt oder mit der Existenz von fehlenden Werten bei der Zielvariablen zusammenhängen, und für die ausreichend gültige Werte innerhalb der fehlenden Fälle der Zielvariablen vorliegen. Um den Einfluss auf das Antwortverhalten zu messen („V“), wird eine dummy Variable gebildet, aus der hervorgeht, ob eine gültige Antwort vorliegt. Diese kann schließlich mit allen Merkmalen aus dem theoretischen Modell korreliert werden. Bei der Berechnung des Anteils brauchbarer Fälle ist darauf zu achten, dass die Grundgesamtheit hier die Gruppe derer ist, für die keine gültige Antwort bei der zu imputierenden Variable vorliegt.

Da mit ICE die Möglichkeit besteht, innerhalb des Imputationsprozesses auch eventuell fehlende Werte der Prädiktor-Variablen zu ersetzen, sollte der eben beschriebene Auswahlprozess für alle verwendeten Prädiktoren wiederholt werden, die soft Missings (s. 4.2) enthalten. Im Hauptmodell können dann ggf. für jede Prädiktor-Variable Untermodelle formuliert werden, anhand derer diese zunächst imputiert werden, bevor sie für die Imputation der eigentlichen Zielvariable herangezogen werden.[12] Da die Vorgehensweise im Einzelnen mit der soeben dargestellten identisch ist, wird auf weitere Ausführungen an dieser Stelle verzichtet.

[12] Um nicht wiederum Untermodelle für die Prädiktoren der Prädiktoren modellieren zu müssen, empfiehlt es sich, lediglich Verknüpfungen zwischen den Variablen der Prädiktoren-Liste zu untersuchen.

4.2. Umgang mit unterschiedlichen Arten von fehlenden Werten

Nicht für alle fehlenden Werte jeder einzelnen verwendeten Variable ist es sinnvoll zu imputieren. So kommt es regelmäßig in Befragungen vor, dass einige Fragen beabsichtigt nicht allen Befragten gestellt werden, da sie im Einzelfall nicht zutreffen. Durch eine entsprechende Filterführung werden dann automatisch fehlende Werte eingesetzt. Eine Imputation ist im Regelfall nur dann erwünscht, wenn eine Antwort verweigert wurde, die Befragten also keine Antwort wussten oder nicht geben wollten.
Im Imputationsmodell müssen daher diese verschiedenen Arten fehlender Werte differenziert werden, sodass nur jene Werte imputiert werden, deren Imputation benötigt wird (so genannte „soft missings"), nicht aber jene, die systembedingt fehlen (so genannte „hard missings"). Diese Differenzierung bezieht sich gleichsam auf alle Variablen, die im Imputationsprozess imputiert werden, also unter Umständen auch auf Prädiktor-Variablen.

Für den Umgang mit hard missings bei ICE wurde bislang noch kein automatischer Programmiervorgang entwickelt, weshalb man derzeit noch etwas aufwändigere Umwege bestreiten muss. Bei STATA 11 und den mi-Kommandos jedoch werden als default nur die soft missings imputiert; hard missings werden bei der Imputation nicht berücksichtigt. Eine nachvollziehbare Methode, mit der sichergestellt werden kann, dass nur die soft missings imputiert werden, ist das so genannte „dummy variable adjustment" (vgl. Allison 2001: 87): Drei Arbeitsschritte sind hierfür notwendig: Zunächst müssen die soft und hard missings identifiziert und definiert werden. Zweitens werden für alle betroffenen Variablen zusätzlich Dummy-Variablen erstellt, aus denen hervorgeht, ob ein hard missing vorliegt. Schließlich werden die hard missings in den ursprünglichen Variablen durch entsprechende repräsentative Werte ersetzt. Sowohl die Ursprungsvariablen (mit den Ersatzwerten für die systembedingten fehlenden Werte) als auch die Dummies sind später Bestandteil des Imputationsmodells. Im Anschluss an die Imputation gilt es lediglich zu bedenken, die Ersatzwerte durch eine Wiederherstellung der hard missings zu löschen.

Beispiel

Für das Haushaltsgesamtvermögen in *ViD* liegen ausschließlich soft missings vor. Da allerdings im Prozess der Imputation auch die übrigen Prädiktor-Variablen mit fehlenden Angaben imputiert werden, kommt dieses Problem der unterschiedlichen Arten fehlender Werte durchaus zum Tragen. Beispielsweise dann, wenn Informationen über den Partner / die Partnerin bei der Imputation berücksichtigt werden sollen. Für diejenigen Befragten, die nicht in einer Partnerschaft leben, liegen bei den Partnerangaben fehlende Werte vor. Diese zu imputieren, wäre nicht nur unnötig, sondern auch falsch. Ein weiteres Beispiel ist das Erwerbseinkommen: Personen, die im Vorjahr nicht erwerbstätig waren und somit kein Einkommen erhalten haben, haben hier korrekterweise keinen gültigen Wert vorliegen.

In Anlehnung an das dummy variable adjustment werden daher die beschriebenen drei Schritte befolgt. Zunächst werden zur Identifikation und Differenzierung der unterschiedlichen missings die soft

missings mit „.“ und die hard missings mit „.a“ oder „.b“ codiert (dies gilt für STATA). Zweitens wird zusätzlich für jede Variable, die über verschiedene missings verfügt, jeweils eine dummy-Variable erstellt, die den Wert 1 annimmt, wenn es sich um ein hard missing handelt. Diese werden später ins Modell aufgenommen. Für den dritten Schritt, in dem die hard missings durch repräsentative Werte ersetzt werden, kommt es darauf an, für jede Variable einen Wert zu finden, der als geeigneter Vertreter funktioniert. Es ist nicht völlig unwesentlich, welche Werte gewählt werden, da sie Einfluss auf die Verteilung nehmen können, welche wiederum für die Imputation herangezogen wird. Dennoch muss der Aufwand für diesen Schritt nicht einer zusätzlichen Imputation nahekommen, da die eingesetzten Werte anschließend wieder gelöscht werden und ihr Einfluss auch während der Imputation durch die Dummies teilweise neutralisiert wird. Im vorliegenden Fall wird den Vorschlägen Allisons (2001) gefolgt, indem für metrische Variablen jeweils der Mittelwert als repräsentativer Wert gewählt wird, und Variablen mit anderen Skalenniveaus für ihre hard missings den Wert 0 erhalten. Beide Variablen-Typen werden anschließend in das Modell aufgenommen. Dadurch, dass für die hard missings ein Wert eingesetzt wurde, werden nur noch die soft missings imputiert. Da jeweils die Dummy-Variable als zusätzlicher Prädiktor im Modell enthalten ist, wird der Einfluss der eingesetzten Werte relativiert.

Diese Vorgehensweise ist zwar umständlich und vergrößert das Imputationsmodell erheblich, der derzeitige Entwicklungsstand der verfügbaren Software lässt allerdings keinen anderen Weg zu, wie auch Diskussionen im offiziellen weltweiten STATA Forum ergeben haben. Das gewählte Vorgehen wirkt sich nicht auf die Verteilung der Schätzwerte aus und entspricht dem aktuellen interdisziplinären Standard.

4.3. Anzahl der Imputationen (*m*)

Eine zentrale Besonderheit und gleichzeitig ein Vorteil der Multiplen Imputation ist, dass die Schätzung der imputierten Werte mehrmals durchgeführt wird und somit Standardfehler angemessen berücksichtigt werden können. Die Anzahl dieser Wiederholungen (*m*) des Imputationsprozesses kann dabei offen gewählt werden. Rein mathematisch gesehen verkleinern sich mit der Anzahl an Wiederholungen automatisch die Standardfehler der geschätzten Parameter, die Schätzung selbst wird präziser. Der zusätzliche Gewinn an Präzision muss allerdings in Relation zu den „Kosten“ betrachtet werden, die durch den Rechenaufwand verursacht werden. *M* unnötig groß zu wählen „is inefficient and computationally expensive with a large dataset“ (Royston/Carlin/White 2009: 254). Dennoch sollte man sich der Tatsache bewusst sein, dass durch die Wahl von *m* Endergebnisse gering abweichen können. Es ist daher ratsam, jeweils die passende Anzahl an Imputationen zu wählen und vor der endgültigen Entscheidung verschiedene Werte ausprobiert, um die angemessene Schwelle für *m* herauszufinden (vgl. ebd.: 258).

Traditionell wurde bisher ein niedriger Wert für *m* gewählt, meist zwischen 3 und 5 (vgl. Royston 2004: 228). Es galt die Annahme, dass die Anzahl der Imputations-Vorgänge derart gering gehalten werden kann, da sich auch bei vielen Wiederholungen die Werte nur unwesentlich ändern. Auch Carlin u.a. gehen davon aus, dass „it is usually sufficient to obtain a relatively small number of imput-

ed datasets, often as few as 3 or 5, because the relative gains in precision from using larger numbers are minor...“ Allerdings ergänzen sie diese Aussage um den Zusatz: „…unless the fraction of missing data is extremely large.“ (Carlin et al. 2003: 228) Wie aus ihren Überlegungen deutlich wird, gilt bei Variablen mit einer hohen Anzahl von fehlenden Werten also ein Spezialfall. In solchen Situationen kann es durchaus sinnvoll sein, *m* zu erhöhen.

In aktuellen Ausführungen wird die Annahme bestätigt, dass *m* eher höher als zu niedrig gewählt werden sollte. Graham, Olchowski und Gilreath (2007) schlagen vor, dass es effizienter ist mehr Imputationen zu nutzen, als zuvor allgemein empfohlen wurde. Sie argumentieren vor allem damit, dass geringe Anzahlen von *m* zu erhöhten Standardfehlern und zu geringerer Power führen (vgl. Graham/Olchowski/Gilreath 2007: 208ff.). Auch sie heben hervor, dass es in zwei Situationen besonders empfehlenswert ist, *m* größer als 5 zu wählen; zum einen bei einer hohen Anzahl an fehlenden Werten, zum anderen bei kleinen bis mittleren Datensätzen. Beide Situationen liegen im hiesigen Fall vor. Zwar schlagen die Autoren vor, *m* in einer Größenordnung zwischen 50 und 100 zu wählen, in Übereinstimmung mit der aufgeführten Literatur kann allerdings davon ausgegangen werden, dass 20 Wiederholungen vollkommen ausreichend und angemessen sind.
Im vorliegenden Imputationsprozess werden daher 20 Imputationen gewählt.

4.4. Kategoriale Variablen

Kategoriale Variablen werden je nach gewähltem Statistikprogramm unterschiedlich im Imputationsprozess berücksichtigt. Generell sollten alle binären Variablen als dummy mit den Werten 0 und 1 codiert werden, damit sie als solche vom Programm erkannt werden. Auch für nominal skalierte Variablen mit mehr als zwei Ausprägungen ist es möglich, diese in einzelne Dummies zu zerlegen. Da einige Programme sie sonst als ordinal behandeln, ist dies mitunter auch angebracht. Allerdings ist es stets vorteilhaft, wenn die Informationen in der Form verarbeitet werden, in der sie vorkommen, das spart Aufwand und beugt Fehlern vor. Mit ICE ist es möglich, die ursprüngliche Struktur der Variablen beizubehalten, das Programm differenziert zwischen den verschiedenen Skalenniveaus.[13] Indem der Vorcode „.o“ in der Variablenliste sowie der Vorcode „.i“ und „.m“ in den jeweiligen Modellen spezifiziert wird (Royston 2009: 472f.), wird gewährleistet, dass kategoriale Variablen mit mehr als 2 Ausprägungen nicht ordinal verwendet werden, sondern jede Kategorie separat in die Analysen einfließt. ICE erstellt automatisch dummy-Variablen für jede Kategorie.
Das Skalenniveau wirkt sich darüber hinaus auch auf die Auswahl der statistischen Verfahren aus, anhand derer die Berechnungen im Imputationsprozess ablaufen. Ähnlich wie das „dummy variable adjustment“ erfordert die Berücksichtigung der Variablenstruktur gesteigerten Aufwand, der in Anbetracht des Mehrgewinns an Präzision jedoch gerechtfertigt sein dürfte.

Die einzelnen getroffenen Vorarbeiten und Entscheidungen münden schließlich in der Formulierung des Modells, also in der Syntax-Programmierung des Imputations-Befehls. Es empfiehlt sich, das Mo-

[13] Weitere Ausführungen zum Umgang mit kategorialen Variablen bei der Multiplen Imputation bei Royston 2009.

dell erst schrittweise aufzubauen, d.h. die einzelnen Variablen nach und nach zu integrieren und auch die einzelnen Optionen erst allmählich aufzunehmen. Auf diese Weise können auftretende Fehlerquellen besser zugeordnet und behoben werden. Auch kann man so die einzelnen Abläufe besser nachvollziehen und vergegenwärtigen.

5. Ergebnis

Nachdem der Imputationsprozess erfolgreich durchlaufen wurde (was mitunter viel Zeit in Anspruch nimmt, da die aufwändigen Berechnungen äußerst viel Arbeitsspeicher benötigen und zudem *m* mal durchgeführt werden müssen), liegen noch immer einige Abschlussarbeiten an, bevor das endgültige Ziel erreicht ist:

- Zunächst ist es unabdingbar, die imputierten Variablen einer ausgiebigen Prüfung zu unterziehen. Es muss nachgehalten werden, ob die eingesetzten Werte plausibel sind. Wie oben erwähnt (vgl. Kap. 3.3), ist es möglich den Analyseschritt sowohl mit mim von ICE, als auch mit den Analyse-Tools von mi durchzuführen. Mit den Befehlen „mi describe“, „mi xeq“ und „mi estimate“ (ggf. herunterladen) kann man sich einen Überblick über die Abläufe und Ergebnisse der Imputation verschaffen. Dazu müssen die imputierten Daten in das mi-Format mit Hilfe des Befehls „mi import“ ICE umgewandelt werden.

  In diesem Schritt ist es vor allem wichtig, die Verteilungen vor und nach der Imputation miteinander abzugleichen. Dabei sollte in erster Linie auf die Standardabweichungen geachtet werden, da sich an ihnen die Güte der Imputation gut ablesen lässt (vgl. etwa Royston 2007 oder Kenward/Carpenter 2007).

  Auch zusätzliche manuelle Tests anhand von Einzelkontrollen sind ratsam, dazu gehört es beispielsweise, Ausreißer auf Plausibilität zu überprüfen oder auch einzelne zufällig ausgewählte Fälle zu kontrollieren. Eine weitere Möglichkeit zur Überprüfung der Modellgüte besteht darin, bereits im Vorfeld einige gültige Werte zu löschen und im Anschluss an die Imputation die neu eingesetzten Werte mit den wahren Werten abzugleichen. Sollten die Tests deutlich Ungereimtheiten identifizieren, ist eine Anpassung des Imputations-Modells ratsam, solange bis die Schätzwerte plausibel sind.
- Nicht vergessen werden sollte auch, die hard missings der Prädiktor-Variablen, welche im Imputationsprozess durch plausible Werte ersetzt wurden, um die unterschiedlichen Arten fehlender Werte zu berücksichtigen, wieder herzustellen.
- Außerdem sollten ggf. die entsprechenden Variablen zurücktransformiert, das heißt „delogarithmiert“ werden, sofern sie aufgrund schiefer Verteilung logarithmiert wurden. Dadurch erhalten die Variablen wieder ihren ursprünglichen Wertebereich. Es gilt zu bedenken, dass unter Umständen noch gerundet werden muss. Ob die Transformation erfolgreich ist, lässt sich durch einen Abgleich einzelner Fälle, die nicht imputiert werden mussten, zwischen den imputierten mit der ursprünglichen Variable kontrollieren.
- Für den Fall, dass wie im angegebenen Beispiel nicht in erster Linie mit allen imputierten Datensätzen gearbeitet werden, sondern ein einzelner Punktschätzer aus den m Werten erstellt werden soll, gilt es diesen noch zu berechnen. Hierfür bietet es sich an, nach Rubin mit der Formel

$\bar{Q} = \frac{1}{m}\sum_{i=1}^{m} \hat{Q}^{(i)}$ der Mittelwert über *m* zu bilden (vgl. Frick/Grabka 2009: 207; Lüdtke et al. 2007: 114; Carlin et al. 2003: 228; Royston 2004: 237). Die entstehende Variable ist es schließlich, die in den Hauptdatensatz aufgenommen und mit der spätere Analysen durchgeführt werden.

- Um auch später noch nachhalten zu können, welche Werte in welcher Form verändert und welche übernommen wurden, sollte zudem eine „Identifikator-Variable" gebildet und im Datensatz aufgenommen werden, die zwischen editierten, imputierten und unveränderten Fällen differenziert.

Beispiel

Im Ergebnis entsteht nach diesen abschließenden Analyseschritten auch für das Anwendungsbeispiel eine einzige Variable, in der für jeden *ViD*-Haushalt eine gültige und plausible Angabe zur Höhe des Haushaltsgesamtvermögens enthalten ist („vermgen_imp").

Tabelle 4: Eckwerte zur Verteilung der endgültigen imputierten Variable

		vermgen (Generierte Variable)	vermgen _imp (nach Imputation)
N	Gültig	335	472
	Fehlend	137	0
Mittelwert		**2.289.866**	**2.448.044**
Standardabweichung		4.036.966	4.308.181
Minimum		200.000	200.000
Maximum		50.000.000	50.300.328
Perzentile	25	750.000	800.000
	50 (Median)	**1.200.000**	**1.500.000**
	75	2.250.000	2.563.579

Die Verteilung der entstandenen Zielvariable zeigt, dass durch die Multiple Imputation das durchschnittliche Vermögen im Vergleich zur generierten Variable vermgen nochmals erhöht wurde (vgl. Tabelle 4). Dies kann einerseits derart gedeutet werden, dass vor allem in Haushalten mit besonders hohen Vermögen häufiger editiert und imputiert werden musste. Andererseits zeigt dieser Einfluss – eigens im Prozess des Editings und der logischen Imputation – auch, dass Vermögenswerte häufig von den Befragten unterschätzt werden, vor allem wenn offene Gesamtbeträge genannt werden sollen. Für Einkommen wurde dieses Phänomen bereits bekundet, auch dieses wird tendenziell in Erhebungen unterschätzt (vgl. Braun/Metzger 2007: 77f.). Dass Korrekturen in solchen Fällen stets in einer Zunahme der Vermögensverteilung resultieren, kann auch in anderen Studien gezeigt werden. So erhöht sich beispielsweise im Sozio-Ökonomischen Panel durch die Editierung und Imputation das durchschnittliche Nettovermögen der Haushalte um knapp 32 Prozent (vgl. Frick/Grabka 2009: 215).
Eine weitere Veränderung der Vermögensverteilung betrifft die Streuung der Vermögenswerte. Durch die Berücksichtigung der Standardfehler (und z.T. durch die Verwendung der logarithmierten Skala) ist diese durch die Imputation nun ebenfalls größer. Das hat zur Folge, dass insgesamt deutlich mehr Werte in der Verteilung vorkommen. Da in Befragungen die Neigung besteht, runde Beträge zu nennen, diese in der Realität jedoch nur selten vorkommen, ist die neue Verteilung dadurch deutlich realistischer geworden. Insgesamt hat sich die Verteilung des Haushaltsgesamtvermögens durch die

Editierung und Imputation jedoch nur unmerklich verändert (siehe Boxplots im Anhang), die statistische Aussagekraft und Datenqualität konnte dagegen merklich erhöht werden.

6. Schlussbemerkungen

Multiple Imputationsverfahren erweisen sich eindeutig als die beste Methode für den Umgang mit fehlenden Werten in komplexen Datensätzen. Die Anzahl verwendbarer Fälle erhöht sich auf ein Maximum, die statistische Aussagekraft der Analysen wird gesteigert und die Informationsfülle aus den Daten ausgeschöpft. Vor allem für Einkommens- und Vermögensangaben, die besonders anfällig für Messfehler, jedoch gleichzeitig besonders entscheidend für sozialwissenschaftliche Analysen sind, bietet es sich an, Imputationen durchzuführen. Zukünftig wird es jedoch nötig sein, die Möglichkeiten der Software einzelner Statistikprogramme auszubauen, vor allem für die Handhabung verschiedener Arten fehlender Werte. Die Bestimmung des Missing-Mechanismus, vor allem aber die Unterscheidung zwischen MAR und MNAR, muss standardisiert werden, um sicherzustellen, dass die Voraussetzungen für eine Imputation gegeben sind. Es drängt sich gerade bei hohen Vermögens- und Einkommensangaben häufig der Verdacht auf, dass die Daten nicht zufällig fehlen. Jedoch fehlen bisher die notwendigen Analysemethoden, um zwischen der Zufälligkeit der fehlenden Daten und der Nicht-Zufälligkeit sicher zu unterscheiden. Eine sensible Analyse der Daten und die logische Herleitung des Missing-Mechanismus müssen deshalb stets Voraussetzung des Imputationsprozesses sein. Auch müssen die Programme insgesamt anwenderfreundlicher werden, um dieses Verfahren einer breiteren Masse an Forschern zugänglich zu machen. Die grundlegenden Herausforderungen für Anwender, etwa die notwendigen Vorbereitungen für eine Multiple Imputation, die Modellformulierung oder die Kontrolle der Ergebnisse können mit einfachen – wenn auch zum Teil aufwändigen – Mitteln bewältigt werden. Der vorliegende Beitrag hat hierfür mögliche Herangehensweisen und Lösungsansätze aufgezeigt.

Es bleibt anzumerken, dass eine Multiple Imputation wie jede Form der Imputation lediglich einen Kompromiss zum Umgang mit Messfehlern darstellt, diese dadurch aber nicht vollständig neutralisiert werden können. Jede Datenveränderung bleibt ein sensibles Unterfangen und mit Ungenauigkeiten behaftet. Kein Datensatz – so auch nicht *ViD* – bietet perfekte Bedingungen, um fehlende Angaben zweifellos treffsicher vorhersagen zu können. Vor allem kommt es daher stets darauf an, sowohl den Prozess der Imputation äußerst sorgfältig zu modellieren und zu überprüfen als auch entsprechend vor- und umsichtig mit den imputierten Datensätzen umzugehen. Selbst die Multiple Imputation als sehr zuverlässiges statistisches Verfahren ist hier in keinem Fall als Allheilmittel zu verstehen: „Multiple imputation is not a panacea. Although it is a powerful and useful tool applicable to many missing data settings, if not used carefully it is potentially dangerous.“ (Horton / Lipsitz 2001: 253)

7. Quellenverzeichnis

Allison, P. D., 2000: Multiple Imputation for Missing Data: A Cautionary Tale. Sociological Methods and Research 28: 301–309.

Allison, P. D., 2001: Missing data. Thousand Oaks, Calif.

Braun, R. und H. Metzger, 2007: Trends in der Entwicklung von Vermögen und Vermögenseinnahmen zukünftiger Rentnergenerationen. Endbericht für das Bundesministerium für Arbeit und Soziales. Online verfügbar unter http://www.bmas.de (Stand: 2008).

Briggs, A., T. Clark, J. Wolstenholme und P. Clarke, 2002): Missing…presumed at random: cost-analysis of incomplete data. Health Economics 12: 377–392.

Carlin, J. B., N. Li, P. Greenwood und C. Coffey, 2003: Tools for analyzing multiple imputed datasets. STATA Journal 3: 226–244.

Collins, L. M., J. Schafer und C. Kam, 2001: A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychological Methods 6: 330–351. (02.06.2010).

Drechsler, J., 2010: Multiple imputation in practice - a case study using a complex German establishment survey. AStA Advances in Statistical Analysis 94/4: 1-26.

Fessler, P., P. Mooslechner und M. Schürz, 2009: Statistische Herausforderungen der Forschung zu Finanzen privater Haushalte im Euroraum. Statistiken 58: H. 1/09: 57-66.

Frick, J. R. und M.M. Grabka, 2007: Editing and Multiple Imputation of Item-Non-Response in the 2002 Wealth Module of the German Socio-Economic Panel (SOEP). Deutsches Institut für Wirtschaftsforschung. Berlin. (Research Notes, 18).

Frick, J. R. und M.M. Grabka, 2009: Erstellung und Analyse einer konsistenten Vermögensverteilungsrechnung für Personen und Haushalte 2002 und 2007 unter Berücksichtigung der personellen Einkommensverteilung. Abschlussbericht. Deutsches Institut für Wirtschaftsforschung. Berlin.

Göthlich, S. E. (2007): Zum Umgang mit fehlenden Daten in großzahligen empirischen Erhebungen. S. 119–134 in: S. Albers, D. Klapper, U. Konradt, A. Walter und J. Wolf (Hg.): Methodik der empirischen Forschung. Wiesbaden: Gabler Verlag.

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

Hohl, K., 2008: Umgang mit fehlenden Werten. Ersetzungmethoden für fehlende Werte kategorialer Variablen in klinischer Datensätze: Vdm Verlag Dr. Müller. http://vts.uni-ulm.de (10.08.2010)

Horton, N. J., S. R. Lipsitz und P. Michael, 2003: A Potential for Bias When Rounding in Multiple Imputation. The American Statistician 57/4: 229–232.

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 61/1: 79–90.

Huster, E., 2009: Reiche und Superreiche in Deutschland - Begriffe und soziale Bewertung. S. 45–53 in: T. Druyen, W. Lauterbach und M. Grundmann (Hg.): Reichtum und Vermögen. Zur gesellschaftlichen Bedeutung der Reichtums- und Vermögensforschung: VS Verlag für Sozialwissenschaften.

Kennickell, A. B., 1998: Multiple Imputation in the Survey of Consumer Finances. http://citeseerx.ist.psu.edu (01.02.2011).

Kenward, M. und J. Carpenter, 2007: Multiple imputation: current perspectives. Statistical methods in Medical Research 16: 199–218.

Kortmann, K. (2011): Vermögen in Deutschland - die methodische Anlage der Untersuchung. S. 15–27 in: W. Lauterbach, T. Druyen und M. Grundmann (Hg.): Vermögen in Deutschland. Heterogenität und Verantwortung. Wiesbaden: VS Verlag für Sozialwissenschaften.

Krug, G., 2010: Fehlende Daten bei der Verknüpfung von Prozess- und Befragungsdaten. Ein empirischer Vergleich ausgewählter Missing Data Verfahren. Methoden — Daten — Analysen 4/1: 27–57.

Lauterbach, W., M. Kramer und M. Ströing, 2011: Vermögen in Deutschland: Konzept und Durchführung. S. 29-53 in: W. Lauterbach, T. Druyen und M. Grundmann (Hg.): Vermögen in Deutschland. Heterogenität und Verantwortung. Wiesbaden: VS Verlag für Sozialwissenschaften.

Lee, K. J., und J.B. Carlin, 2010: Multiple Imputation for Missing Data: Fully Conditional Specification Versus Multivariate Normal Imputation. American Journal of Epidemiology 171/5: 624–632.

Lüdtke, O., A. Robitzsch, U. Trautwein und O. Köller, 2007: Umgang mit fehlenden Werten in der psychologischen Forschung. Problem und Lösungen. Psychologische Rundschau 58/2: 103–117.

Marchenko, Y.V. und J.P. Reiter, 2009: Improved degress of freedom for multivariate significance test obtained from multiply imputed, small-sample data. STATA Journal 9/3: 388–397.

McKnight, P., K. McKnight, S. Sidani und A. J. Figueredo, 2007: Missing data. A gentle introduction. New York: Guilford Press.

Rässler, S., D.B. Rubin und E.R. Zell, 2007: Incomplete Data in Epidemiology and Medical Statistics. S. 569–601 in: C.R. Rao, J. P. Miller und D. C. Rao (Hg.): Handbook of Statistics : Epidemiology and Medical Statistics: Elsevier, Volume 27.

Riphahn, R. T. und O. Serfling, 2002: Item Non-Response on Income and Wealth Questions. IZA Discussion Papers 573: 1–37.

Royston, P., 2004: Multiple imputation of missing values. STATA Journal 4/3: 2227–2241.

Royston, P., 2005a: Multiple Imputation of missing values: update. STATA Journal 5/2: 188–201.

Royston, P., 2005b: Multiple Imputation of missing values: Update of ice. STATA Journal 5/4: 527–536.

Royston, P., 2007: Multiple imputation of missing values: further update of ice, with an emphasis on interval censoring. STATA Journal 7/4: 445–464.

Royston, P., J.B. Carlin und I.R. White, 2009: Multiple Imputation of missing values: New features for mim. STATA Journal 9/2: 252–264.

Rubin, D.B., 2004: Multiple imputation for nonresponse in surveys. New York: Wiley.

Schafer, J.L., 1997: Analysis of incomplete multivariate data. Boca Raton: Chapman & Hall.

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

Spiess, M. 2010: Der Umgang mit fehlenden Werten. S. 117–142 in: C. Wolf, und H. Best (Hg.): Handbuch der sozialwissenschaftlichen Datenanalyse. Wiesbaden: VS Verlag für Sozialwissenschaften.

Spiess, M. und J. Goebel, 2005: on the effect of item nonresponse on the estimation of a two-panel-waves wage equation. Allgemeines Statistisches Archiv 89: 63–74.

STATA Press (Hg.), 2009: STATA multiple-imputation reference manual. Release 11. College Station, Tex.: STATA Press.

van Buuren, S., H.C. Boshuizen und D.L. Knook, 1999: Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis. Statistics in Medicine 18: 681–694.

van Buuren, S., 2007: Multiple imputation of discrete and continuous data by fully conditional specification. Statistical methods in Medical Research 16: 219–242.

ANHANG

Fragebogenausschnitte - Vermögensangaben

3.5 Wenn Sie die von Ihnen genannten Geldanlagen zusammenrechnen, wie hoch ist das Geldvermögen Ihres Haushalts insgesamt?

In Euro	*Gesamtwert der Geldanlagen des Haushalts*
Weniger als 100.000	☐
100.000 bis unter 250.000	☐
250.000 bis unter 500.000	☐
500.000 bis unter 1 Million	☐
1 Million bis unter 2 Millionen	☐
2 Millionen bis unter 5 Millionen	☐
5 Millionen und mehr	☐
Trifft nicht zu	☐
Keine Angabe	☐

3.9 Wenn Sie den heutigen Verkaufswert der Vermögensbestände zusammenrechnen, also den Wert aller Haushaltsvermögen, die Sie nicht bereits zu den Geldanlagen gezählt haben, in welche Gruppe wäre dieser einzuordnen?

In Euro	*Verkaufswert aller sonstigen Vermögenswerte des Haushalts*
Weniger als 250.000	☐
250.000 bis unter 500.000	☐
500.000 bis unter 1 Million	☐
1 Million bis unter 2 Millionen	☐
2 Millionen bis unter 5 Millionen	☐
5 Millionen bis unter 10 Millionen	☐
10 Millionen und mehr	
Trifft nicht zu	☐
Keine Angabe	☐

3.10 Wir bitten Sie noch einmal, alle Vermögensbestände Ihres Haushalts zusammenzurechnen und uns den Gesamtwert des Haushaltsvermögens möglichst genau zu nennen. Auf welchen Betrag kommen Sie ungefähr?

_____ Euro Gesamtvermögen des Haushalts ☐ *Keine Angabe*

Boxplots für das generierte Vermögen und das endgültige imputierte Vermögen