Ilmenauer Beiträge zur Wirtschaftsinformatik

Herausgegeben von U. Bankhofer, V. Nissen
D. Stelzer und S. Straßburger

Tobias Rockel

Gütevergleich von Imputationsverfahren – Eine Analyse existierender Simulationsstudien

Arbeitsbericht Nr. 2017-01, April 2017

Technische Universität Ilmenau
Fakultät für Wirtschaftswissenschaften
Institut für Wirtschaftsinformatik

Autor: Tobias Rockel

Titel: Gütevergleich von Imputationsverfahren – Eine Analyse existierender Simulationsstudien

Ilmenauer Beiträge zur Wirtschaftsinformatik Nr. 2017-01, Technische Universität Ilmenau, 2017

ISSN 1861-9223

ISBN 978-3-938940-59-4

urn:nbn:de:gbv:ilm1-2017200274

Anschrift: Technische Universität Ilmenau, Fakultät für Wirtschaftswissenschaften und Medien, Institut für Wirtschaftsinformatik, PF 100565, D-98684 Ilmenau.
http://www.tu-ilmenau.de/wid/forschung/ilmenauer-beitraege-zur-wirtschaftsinformatik/

Gliederung

Zusammenfassung: Das vorliegende Arbeitspapier aggregiert die Erkenntnisse aus 125 Simulationsstudien, die Imputationsverfahren vergleichen. Dazu werden zunächst der Aufbau der Studien untersucht und die Studien mit verlässlichen Ergebnissen ausgewählt. Diese Studien bilden die Basis für eine Analyse der Imputationsverfahren. Hierbei werden die Verfahren zunächst separat betrachtet und danach paarweise miteinander verglichen. Zusammenfassend ergeben beide Untersuchungen, dass die Imputation mittels adaptiver Regression, die multiple Imputation und die ML-Parameterschätzverfahren am besten zur Behandlung fehlender Werte geeignet sind. Über den Verfahrensvergleich hinaus erlauben die Studien auch Rückschlüsse über Faktoren, die die Qualität der Imputation beeinflussen. Die Studien zeigen, dass sowohl eine größere Anzahl an Objekten als auch ein geringere Anteil fehlender Werte zu besseren Ergebnissen führen. Die Aggregation der Studien zeigt auch weiteren Forschungsbedarf auf. Zum einen sind die Auswirkungen der Merkmale auf die Imputationsqualität nicht eindeutig und zum anderen sind viele Verfahren noch nie oder nicht häufig genug für belastbare Aussagen miteinander verglichen worden. Insbesondere wurden die drei besten Verfahren in keiner Studie direkt miteinander verglichen.

Schlüsselworte: Imputation, Simulationsstudien, fehlende Werte, missing data

1 Problemstellung

Die meisten Verfahren zur Datenanalyse gehen davon aus, dass eine vollständige Datenmatrix vorliegt (vgl. Schafer und Graham 2002, S. 147). Jedoch enthalten in der Realität viele Datenmatrizen fehlende Werte. So stellen z. B. Eekhout et al. (2012, S. 729–731) bei einer Untersuchung von 285 epidemiologischen Studien fest, dass bei mindestens 262 Studien fehlende Werte aufgetreten sind. Backhaus und Blechschmidt (2009, S. 266) behaupten sogar, dass in der Realität praktisch keine Datenmatrix ohne fehlende Werte existiert.

Hierdurch ist die direkte Analyse von Datenmatrizen mittels herkömmlicher Verfahren meist nicht möglich. Vielmehr muss zunächst eine Strategie zum Umgang mit den fehlenden Werten festgelegt werden. Dafür stehen fünf verschiedene Ansätze zur Verfügung: Die Eliminierung unvollständiger Objekte oder Merkmale, die Imputation der fehlenden Werte, die Schätzung von Parametern anhand der unvollständigen Datenmatrix, die Anpassung multivariater Verfahren, sodass Datenmatrizen mit fehlenden Werten direkt analysiert werden können, und Sensitivitätsanalysen. Eine ausführliche Darstellung dieser fünf Ansätze ist bei Bankhofer (1995) zu finden.

Die Wahl eines geeigneten Verfahrens ist unter anderem vom vorliegenden Ausfallmechanismus abhängig. Die Ausfallmechanismen beschreiben den stochastischen Zusammenhang zwischen der Missing Data (MD) Indikatormatrix, welche anzeigt, ob ein Wert beobachtet ist oder nicht, und der Datenmatrix. Falls die MD-Indikatormatrix (stochastisch) unabhängig von der Datenmatrix ist, dann werden die Daten als Missing Completely at Random (MCAR) bezeichnet. Dies stellt einen Spezialfall von Missing at Random (MAR) dar. Denn MAR erlaubt, dass die Wahrscheinlichkeit für das Fehlen von Werten von den beobachteten – aber nicht von den unbeobachteten Werten – abhängt. Wenn das Fehlen der Werte auch von den unbeobachteten Werten abhängt, liegt ein Not Missing at Random (NMAR) Mechanismus vor (vgl. Little und Rubin 2002, S. 12).

Im Folgenden wird insbesondere auf Imputationsverfahren genauer eingegangen. Mit Hilfe von Imputationsverfahren werden fehlende Werte durch einen geschätzten Wert ersetzt. Nachdem die Imputation abgeschlossen ist, steht folglich eine vervollständigte Datenmatrix zur Verfügung, die mit herkömmlichen Analyseverfahren untersucht werden kann. Daher besteht in diesem Fall eine Abhängigkeit zwischen der Imputation und der Ergeb-

nisqualität der folgenden Datenanalyse. Aus diesem Grund ist es eine wichtige Frage, welches Imputationsverfahren zur Ersetzung von fehlenden Werten verwendet werden sollte.

Zur Untersuchung dieser Frage werden unter anderem Simulationsstudien eingesetzt. Da mittlerweile eine Vielzahl dieser Studien existiert, werden in der Literatur immer wieder Zusammenfassungen einzelner Studien gegeben (vgl. Raymond 1986, S. 408–410, Roth 1994, S. 540–545, Tsikriktsis 2005, S. 57–58, Aittokallio 2010, S. 257–259, Devi Priya und Sivaraj 2015, S. 67–68). Diese Zusammenfassung geschieht bei den genannten Autoren entweder in Form von Fließtext oder die Studien werden einzeln in Form von Tabellen dargestellt. Beide Darstellungsformen erlauben eine übersichtliche Wiedergabe von Studiendetails, aber limitieren gleichzeitig die Anzahl an untersuchbaren Studien. So basiert jede der genannten Zusammenfassungen auf weniger als 30 Studien.

Für das vorliegende Arbeitspapier wurde eine umfangreiche Literaturrecherche durchgeführt. Dabei wurde nach Studien gesucht, die Imputationsverfahren vergleichen. In die folgende Untersuchung wurden alle Ergebnisse eingeschlossen, die mindestens zwei Imputationsverfahren vergleichen und sich nicht ausschließlich auf longitudinale Datenmatrizen und Methoden für longitudinale Daten beschränken.[1] Mit Hilfe dieser Kriterien ließen sich 125 Studien identifizieren.

Diese 125 Studien stellen die Basis für das Arbeitspapier dar. Auf Grund der großen Anzahl an Studien wird von einer Darstellung jeder einzelnen Studie Abstand genommen. Stattdessen wird zunächst der Aufbau aller Studien im zweiten Kapitel gemeinsam analysiert. Hierbei zeigt sich, dass über die Hälfte der Studien nicht auf Imputationsverfahren beschränkt sind, sondern darüber hinaus andere MD-Verfahren miteinbeziehen. Daher geht die folgende Analyse in Teilen auch allgemein auf MD-Verfahren ein. Im dritten Kapitel werden die MD-Verfahren anhand der untersuchten Studien verglichen. Ferner werden die Auswirkungen von verschiedenen Faktoren auf die Güte der Verfahren untersucht. Am Ende der Arbeit werden die wichtigsten Erkenntnisse zusammengefasst und Forschungslücken aufgezeigt.

[1] Longitudinale Daten wurden ausgeschlossen, da für diese Datenstruktur spezielle Imputationsverfahren existieren.

2 Aufbau der untersuchten Simulationsstudien

Bei den untersuchten Simulationen gibt es zwei verschiedene Vorgehensweisen, die sich im Aufbau der Studien widerspiegeln. Bei der ersten Vorgehensweise verwenden die Autoren zur Untersuchung der Imputationsverfahren reale Datenmatrizen mit a-priori fehlenden Werten, auf welche sie die Imputationsverfahren anwenden. Hingegen wird bei der zweiten Vorgehensweise eine Datenmatrix ohne fehlende Werte zugrunde gelegt. Aus dieser werden zunächst Werte gelöscht, bevor die Imputationsverfahren angewendet werden.

Abbildung 1: Vorgehensweise bei einer Datenmatrix mit fehlenden Werten

Der Aufbau einer Studie bei der Verwendung von Datenmatrizen mit fehlenden Werten ist schematisch in der Abbildung 1 gezeigt. Eine solche Studie besteht meist aus drei Schritten. Im Ersten werden eine oder mehrere reale Datenmatrizen mit fehlenden Werten ausgewählt. Anschließend wird jedes Imputationsverfahren mindestens einmal auf jede Datenmatrix angewendet. Dann werden die Ergebnisse der verschiedenen Verfahren miteinander verglichen (vgl. z. B., Myrtveit et al. 2001; Pérez et al. 2002; Barzi und Woodward 2004; de Souto et al. 2015).

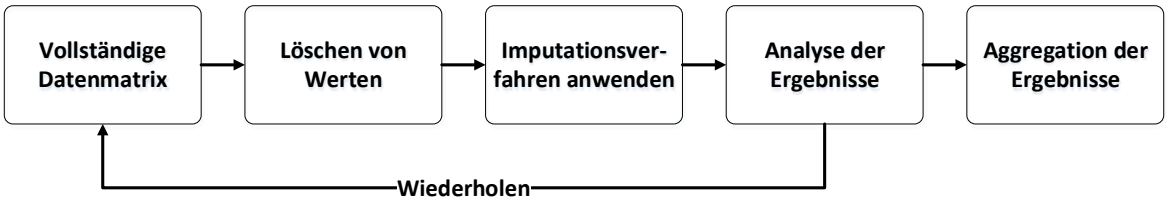

Abbildung 2: Vorgehensweise bei einer vollständigen Datenmatrix

Bei der zweiten Vorgehensweise, dargestellt in der Abbildung 2, werden als Datengrundlage entweder reale, vollständige Datenmatrizen[2] oder synthetische Datenmatrizen verwendet. Aus diesen werden dann Werte gelöscht, um so eine Datenmatrix mit fehlenden Werten zu erzeugen. Anschließend werden die Imputationsverfahren angewendet und die Ergebnisse der Verfahren analysiert. Dies geschieht durch einen Vergleich mit Ergebnissen

[2] Ein Teil der Studien verwendet neben realen Datenmatrizen, die a-priori keine fehlenden Werte besitzen, auch Datenmatrizen, die ursprünglich fehlende Werte enthielten. In diesem Fall werden die Datenmatrizen vorverarbeitet, so dass sie am Ende keine fehlenden Werte mehr enthalten. Dies geschieht in der Regel durch den Ausschluss von Objekten (oder Merkmalen) mit fehlenden Werten (vgl. z. B. Branden und Verboven 2009, Ali et al. 2011, Hallgren und Witkiewitz 2013). Ein Sonderfall stellt die Studie von Ambler et al. 2007 dar, in der die fehlenden Werte einmal imputiert werden. Diese imputierte Datenmatrix wird dann als „vollständige“ Datenmatrix verwendet.

basierend auf der vollständigen Datenmatrix (vgl. z. B. Bello 1993, Batista und Monard 2003, Farhangfar et al. 2008) oder durch einen Vergleich mit bekannten Simulationsparametern (vgl. z. B. Schafer und Graham 2002, Demirtas und Hedeker 2008, Groenwold et al. 2012). Bei über 80 % der Studien wird der Vorgang der Datenerzeugung (falls die Daten simuliert sind), der Löschung der Werte und der Anwendung der Imputationsverfahren sowie die anschließende Analyse mehrmals wiederholt. Zum Abschluss werden die Ergebnisse der Wiederholungen aggregiert. Details hierzu werden im Abschnitt 2.4 dargestellt. Die einzelnen Schritte der Studien werden in den folgenden Unterkapiteln genauer untersucht.

2.1 Verwendete Datenmatrizen

Wie bereits zuvor erwähnt, lassen sich die verwendeten Datenmatrizen in vollständigen und unvollständigen sowie in simulierte und reale unterteilen. Eine Übersicht über die verwendeten Datenmatrizen gibt die Tabelle 1.[3] Von den 125 untersuchten Vergleichsstudien verwenden 29 Datenmatrizen mit fehlenden Werten. Ferner vergleichen 102 Studien die Imputationsverfahren anhand vollständiger Datenmatrizen. Sechs Studien verwenden sowohl Datenmatrizen mit fehlenden Werten als auch vollständige Datenmatrizen und werden in der Tabelle doppelt erfasst. Daher summieren sich 102 und 29 nicht zu 125. Bei diesen sechs Studien wird eine umfangreiche Untersuchung anhand vollständiger Datenmatrizen vorgenommen. Gleichzeitig werden die auf den vollständigen Datenmatrizen basierenden Ergebnisse mit einem bzw. mehreren Datenmatrizen mit fehlenden Werten verglichen (vgl. Feelders 1999, Acuna und Rodriguez 2004, Ibrahim et al. 2005, Liu et al. 2005, García-Laencina et al. 2010, Luengo et al. 2010).

	unvollständig	vollständig	gesamt
real	29	66	91
simuliert	0	44	44
gesamt	29	102	

Tabelle 1: Verwendete Datenmatrizen

Alle Datenmatrizen mit a-priori fehlenden Werten sind reale Datenmatrizen. Es ist zwar theoretisch möglich, unvollständige Datenmatrizen direkt zu simulieren, jedoch wird dies in den untersuchten Studien nicht getan. Ein Grund hierfür ist, dass eine direkte Simulation

[3] Die Randhäufigkeiten in Tabelle 1 entsprechen teilweise nicht den Zeilen- bzw. Spaltensummen, da ein Teil der Autoren verschiedene Kombinationen verwendet. Solche Studien werden in der Tabelle mehrfach erfasst.

von unvollständigen Datenmatrizen – insbesondere bei komplexen Ausfallmechanismen und Ausfallmustern – wesentlich aufwendiger ist als die Erzeugung einer zunächst vollständigen Datenmatrix und anschließendes Löschen von Werten. Daher basieren alle 29 Studien mit unvollständigen Datenmatrizen auf realen Datenmatrizen mit a-priori fehlenden Werten.

Außerdem verwenden 66 Studien reale vollständige Datenmatrizen, wobei vier der 66 Studien sowohl reale vollständige als auch reale unvollständige Datenmatrizen heranziehen (vgl. Acuna und Rodriguez 2004, Liu et al. 2005, García-Laencina et al. 2010; Luengo et al. 2010). Insgesamt werden von 91 der 125 Studien reale Datenmatrizen zum Vergleich der Imputationsverfahren herangezogen. Im Vergleich dazu verwenden 44 der Studien simulierte Datenmatrizen. In zehn Studien werden dabei sowohl simulierte als auch reale Datenmatrizen verwendet.

2.2 Fehlende Werte

Unabhängig davon, ob die fehlenden Werte erzeugt werden oder bereits vorliegen, stehen bei ihrer Betrachtung die drei folgenden Einflussfaktoren im Mittelpunkt:

- Ausfallmuster
- Ausfallmechanismus
- Anteil fehlender Werte

In der Tabelle 2 werden die von den Studien verwendeten Kombinationen von Ausfallmechanismen und -mustern dargestellt. Das Ausfallmuster beschreibt, ob fehlende Werte nur in einem oder in mehreren Merkmalen auftreten. Im ersten Fall wird das Ausfallmuster als univariat und im zweiten Fall als multivariat bezeichnet (vgl. van Buuren 2012, S. 95).

Datenmatrizen mit real fehlenden Werten werden in der Tabelle 2 unter dem Ausfallmechanismus „real“ aufgeführt, da der zugrundeliegende Ausfallmechanismus in diesem Fall normalerweise nicht bekannt ist und ohne Nacherhebung der fehlenden Werte auch nicht zuverlässig ermittelt werden kann (vgl. Schafer und Graham 2002, S. 152).[4] Ferner werden Studien, die verschiedene Kombinationen von Ausfallmechanismen und -mustern verwenden, mehrfach erfasst.

[4] Ein Sonderfall ist die Arbeit von Cox und Folsom (1978). Cox und Folsom verwenden eine reale Datenmatrix mit fehlenden Werte, den sie durch Nacherhebung vervollständigen. Diese Datenmatrix wird daher im vorherigen Kapitel unter die vollständigen realen Datenmatrizen eingeordnet und in diesem Kapitel unter dem Ausfallmechanismus „real“ einsortiert. Alle anderen Datenmatrizen mit „realem“ Ausfallmechanismus werden im vorherigen Kapitel zu den unvollständigen Datenmatrizen gezählt, da alle anderen Studien keine Nacherhebung durchführen.

	univariat	multivariat	gesamt
MCAR	19	69	88
MAR	14	24	37
NMAR	9	13	22
real	4	30	30
gesamt	29	100	

Tabelle 2: Verwendete Ausfallmechanismen und -muster

Aus der Tabelle 2 ist ersichtlich, dass über 70 % der Studien einen MCAR-Ausfallmechanismus, überwiegend in Verbindung mit einem multivariaten Ausfallmuster, verwenden. Im Gegensatz dazu wird der Ausfallmechanismus NMAR selten untersucht. Ein Grund hierfür könnte die einfache Umsetzbarkeit von MCAR-Ausfallmechanismen im Rahmen von Simulationen sein. Hingegen gibt es für die Umsetzung von MAR und NMAR wesentlich mehr Möglichkeiten, wodurch die Auswahl einer konkreten Spezifikation erschwert wird. Daher unterscheiden sich die in den Studien verwendeten Spezifikationen von MAR und NMAR wesentlich stärker als dies bei MCAR der Fall ist.

Weiterhin fällt in der Tabelle 2 auf, dass bei real fehlenden Werten das Verhältnis von multivariat zu univariat deutlich größer ist als bei den simulierten Ausfallmechanismen. Dies liegt zum einen daran, dass in der Realität fast immer ein multivariates Muster vorliegt, und zum anderen die Simulation multivariater Muster komplexer ist als die Simulation eines univariaten Musters. Diese höhere Komplexität trifft besonders für die Simulation von MAR und NMAR zu. Hingegen spielt sie bei MCAR eine eher untergeordnete Rolle. Dies spiegelt sich in den Ergebnissen der Tabelle 2 wider, da das Verhältnis von multivariat zu univariat bei MAR und NMAR deutlich kleiner ist als bei MCAR.

Neben dem Ausfallmechanismus und dem Ausfallmuster ist der dritte wichtige Einflussfaktor der Anteil fehlender Werte. Eine Variation des Anteils fehlender Werte ist bei der Verwendung eines „realen“ Ausfallmechanismus entweder durch die Betrachtung unterschiedlicher Datenmatrizen oder durch mehrfache Erhebung ähnlicher Daten möglich. Der zweite Weg ist sehr teuer und wird daher in keiner der untersuchten Studien gewählt. Aber in 13 der 30 Studien mit „realem“ Ausfallmechanismus werden die Datenmatrizen und damit auch der Anteil fehlender Werte variiert.

Bei der Verwendung vollständiger Datenmatrizen ist die Änderung der Ausfallrate zumindest bei der Simulation eines MCAR-Ausfallmechanismus – dem am häufigsten verwendeten Ausfallmechanismus – wesentlich einfacher. Hierfür muss normalerweise nur ein Pa-

rameter in der Simulation angepasst werden. Daher variieren über 75 % der Studien, die vollständige Datenmatrizen verwenden, den Anteil fehlender Werte, wie aus der Abbildung 3 hervorgeht.

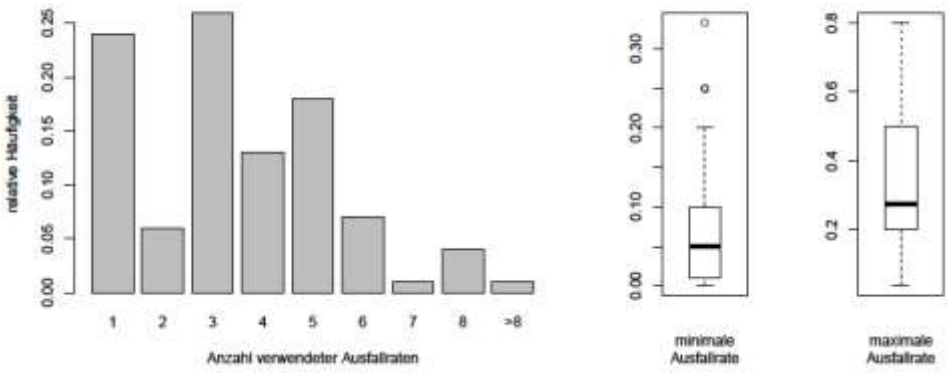

Abbildung 3: Anzahl verwendeter Ausfallraten bei der Simulation fehlender Werte sowie minimale und maximale Ausfallrate, bei Variation des Anteils fehlender Werte

Falls der Anteil fehlender Werte variiert wird, geschieht dies größtenteils auf zwei bis sechs Stufen. Dabei wählen die Studienautoren in über 90 % der Fälle eine minimale Ausfallrate von 10 % oder weniger. Als Startwerte werden in der Regel 1 % fehlende Werte (21 % der Studien), 5 % fehlende Werte (31 % der Studien) oder 10 % fehlende Werte (23 % der Studien) verwendet. Als Endpunkte dienen häufig 20 % fehlende Werte (31 % der Studien) oder 50 % fehlende Werte (15 % der Studien). Insgesamt beschränken sich über die Hälfte der Studien auf weniger als 30 % fehlende Werte und über 80 % der Studien simulieren nicht mehr als 50 % fehlende Werte. Den höchsten Anteil fehlender Werte in den untersuchten Studien simulieren Penone et al. (2014) mit einer Ausfallrate von 80 %.

Theoretisch können innerhalb einer Simulation sowohl der Ausfallmechanismus, die Ausfallrate als auch das Ausfallmuster variiert werden. So wird auch in 71 % aller Studien die Ausfallrate verändert. Hingegen modifizieren nur 28 % der Studienautoren den Ausfallmechanismus und das Ausfallmuster wird in fast allen Studien konstant gehalten. Unter der Annahme, dass sich die Autoren auf die Variation der in ihren Augen wichtigsten Faktoren beschränken, lässt sich ableiten, dass von den drei genannten die Ausfallrate als die wichtigste Einflussquelle auf die Bewertung der Imputationsverfahren angesehen wird.

2.3 Untersuchte Imputationsverfahren

Der zentrale Aspekt jeder Studie sind die untersuchten Imputationsverfahren. Neben den Imputationsverfahren werden in den Studien zum Teil noch weitere MD-Verfahren miteinbezogen. Eine Übersicht über die Verfahren und ihre Verwendungshäufigkeit gibt die Abbildung 4. Dort ist jedes MD-Verfahren aufgeführt, das mindestens fünfmal verwendet wird. Verfahren, die seltener untersucht werden, sind in der Regel sehr speziell oder basieren auf Ideen, die in der Literatur nicht weiterverfolgt wurden. Sie sind in der Abbildung unter dem Begriff „sonstige“ zusammengefasst.

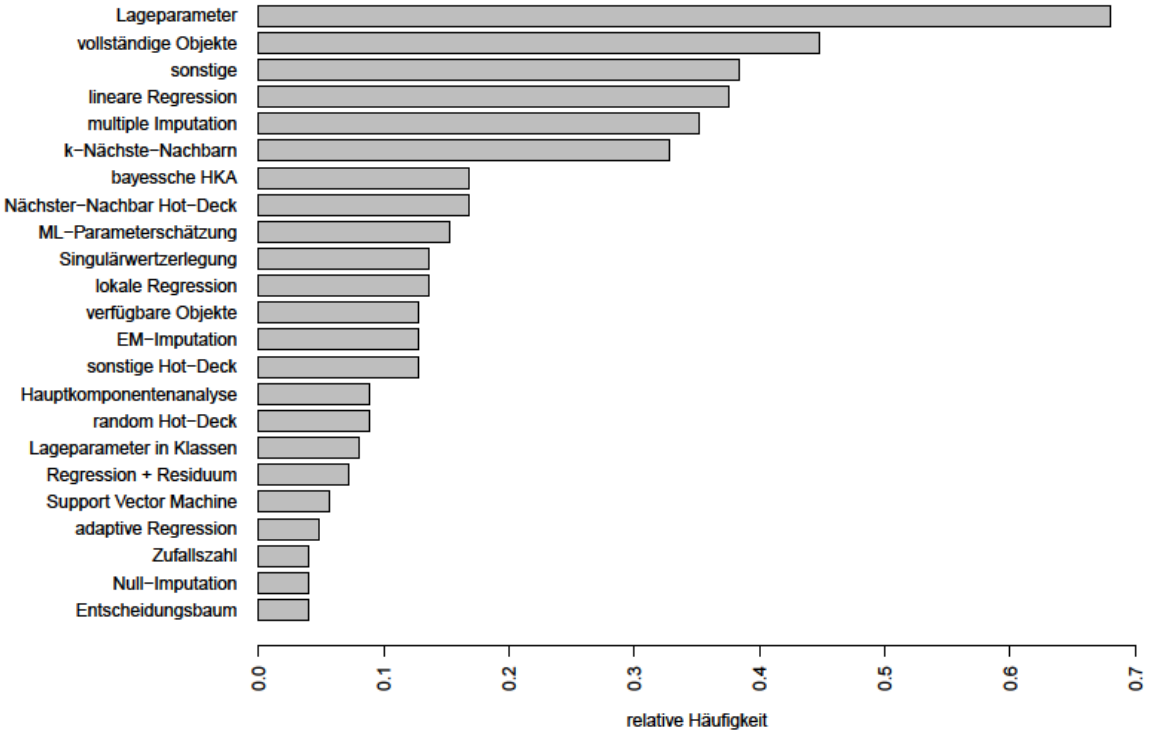

Abbildung 4: Untersuchungshäufigkeit verschiedener MD-Verfahren

Die Abbildung 4 zeigt, dass die Imputation eines Lageparameters am häufigsten untersucht wird. Sie wird in über zwei Dritteln der Studien einbezogen. Mit deutlichen Abstand folgt die Analyse der vollständigen Objekte, die in 45 % der Studien eingeschlossen wird. Bei ihr handelt es sich nicht um ein Imputationsverfahren sondern um ein Eliminierungsverfahren. Sowohl die Imputation eines Lageparameters als auch die Analyse der vollständigen Objekten sind sehr einfach Verfahren und insbesondere die Verwendung der vollständigen Objekte ist regelmäßig als Standard in datenanalystischen Softwarepakten eingestellt (vgl. van Buuren 2012, S. 4). Beide Verfahren werden in den Studien daher oft als Vergleichsbasis herangezogen, wodurch sich ihre häufige Verwendung erklären lässt.

Neben diesen beiden Verfahren werden in etwa einem Drittel der Studien die Imputation mittels linearer Regression, multiple Imputation und die k-Nächste-Nachbarn-Imputation einbezogen. Die Gründe für die vergleichsweise häufige Verwendung dieser drei Verfahren sind unterschiedlich. So zählt die Imputation mittels linearer Regression zu den verbreitetsten Imputationsmethoden in der MD-Literatur (vgl. Bankhofer 1995, S. 126). Ferner wird die multiple Imputation unter anderem von Rubin (1996, S. 473), Allison (2001, S. 2), Schafer und Graham (2002, S. 147) sowie Enders (2010, S. 37) als eine Standardmethode zum Umgang mit fehlenden Daten empfohlen. Hingegen dient die k-Nächste-Nachbarn-Imputation vor allem in neueren Studien – ähnlich wie die Imputation eines Lageparameters oder die Analyse der vollständigen Objekte – als eine Vergleichsbasis für andere Verfahren (vgl. Aittokallio 2010, S. 255).

Alle anderen Verfahren werden in weniger als 20 % der Studien untersucht und damit deutlich seltener als die fünf vorher genannten. Dies ist auch am Sprung zwischen der Imputation mittels k-Nächster-Nachbarn und bayesscher Hauptkomponentenanalyse in der Abbildung 4 zu erkennen. Insgesamt geht aus der Abbildung 4 hervor, dass die Auswahl der herangezogenen MD-Verfahren relativ heterogen ist.

2.4 Analyse der Ergebnisse

Bei der Analyse der Ergebnisse existieren zwischen den beiden Vorgehensweisen unterschiedliche Herangehensweisen, um die Güte der verwendeten Imputationsverfahren zu beurteilen. Bei der Verwendung von vollständigen Datenmatrizen werden die Resultate der Imputationsverfahren entweder mit Ergebnissen basierend auf der vollständigen Datenmatrix oder mit bekannten Simulationsparametern verglichen. Häufig werden die Abweichungen zwischen diesen „optimalen“ Ergebnissen und den Ergebnissen anhand der Imputationsverfahren ermittelt. Je geringer diese Abweichungen sind, desto besser wird das Verfahren beurteilt (vgl. z. B. Roth et al. 1999, Schafer und Graham 2002, Groenwold et al. 2012).

Da bei der Verwendung von unvollständigen Datenmatrizen weder die wahren Parameter bekannt sind noch eine vollständige Datenmatrix zum Vergleich vorliegt, werden die Ergebnisse der Imputationsverfahren meist untereinander verglichen (vgl. z. B. Crawford et al. 1995, Barzi und Woodward 2004, Saunders et al. 2006). Dieses Vorgehen erlaubt nur relative Aussagen wie beispielsweise: „Verfahren A schätzt die Varianz von Variable X kleiner als Verfahren B“. Allerdings lässt sich in der Regel aus solchen Aussagen nicht ab-

leiten, welches der beiden Verfahren eine bessere Schätzung der Varianz liefert. Daher kann bei der Verwendung von unvollständigen Datenmatrizen normalerweise nicht beurteilt werden, welches Verfahren zu besseren Ergebnissen führt.

Neben diesen grundsätzlichen Unterschieden der beiden Vorgehensweisen werden in den Studien auch verschiedene Gütekriterien verwendet, anhand derer die Verfahren beurteilt werden. Dabei lassen sich die Kriterien grob in drei Gruppen einteilen. Die erste Gruppe untersucht die Auswirkungen der Imputationsverfahren auf einzelne Werte, die zweite auf einzelne Parameterschätzungen und die dritte Auswirkungen auf ganze Modelle (z. B. auf Regressionsmodelle oder Entscheidungsbäume).

In der Abbildung 5 ist dargestellt, wie häufig verschiedene Gütekriterien in den untersuchten Studien verwendet werden. In der Abbildung wird jedes Kriterium, das in mindestens fünf Studien angewendet wird, einzeln aufgeführt. Seltener verwendete Kriterien werden in Gruppen zusammengefasst.[5]

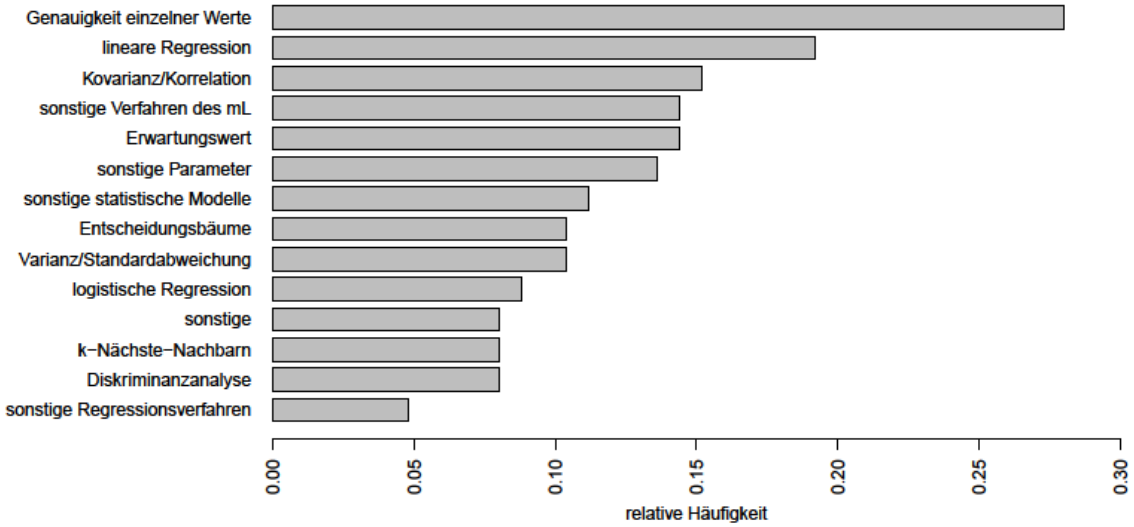

Abbildung 5: Häufigkeit verwendeter Vergleichskriterien

Am häufigsten werden die Imputationsverfahren in den Studien anhand der Genauigkeit einzelner Werte beurteilt. Die Operationalisierung dieses Kriteriums ist in den Studien nicht einheitlich. So wird zur Messung in verschiedenen Studien z. B. der mittlere quadra-

[5] Unter den sonstigen Verfahren des maschinellen Lernens (mL) befinden sich Neuronale Netze, Support Vector Machines, Naive Bayes und weitere klassische Verfahren aus dem Bereich des maschinellen Lernens. Der Teil sonstige Parameter umfasst u. a. Auswirkungen auf Cronbachs Alpha, Quantile, multiple Korrelation, t-Statistik sowie Momentenkoeffizienten. Zu den sonstigen statistischen Modellen gehören vor allem die Hauptkomponentenanalyse, Faktoranalyse, Conjoint Analyse, Varianzanalyse und k-Means. Weitere Verfahren, die keiner der vorherigen Gruppen zugeordnet werden können, sind u. a. die Dauer zum Erlernen der Verfahren, die benötigte Rechenzeit der Verfahren und spezielle Bewertungskriterien aus einzelnen Fachdisziplinen. Diese werden unter dem Punkt „sonstige“ zusammengefasst.

tische Fehler oder die mittlere absolute Abweichung verwendet. Aber auch die Korrelation zwischen imputierten und richtigen Werten findet Anwendung. Die Unterschiede zwischen den Operationalisierungen sind nicht nur theoretischer Natur, sondern können auch zu unterschiedlichen Beurteilungen der untersuchten Verfahren führen (vgl. Oh et al. 2011, S. 84).

Unabhängig von der konkreten Ausgestaltung setzt jede Untersuchung der Genauigkeit eine vollständige Datenmatrix voraus, mit dem die imputierten Werte verglichen werden. Außerdem lassen sich mit diesem Kriterium normalerweise nur Imputationsverfahren vergleichen, da weder Parameterschätzverfahren noch Eliminierungsverfahren eine „Vorhersage“ der fehlenden Werte liefern.[6] Diese beiden Restriktionen beschränken die Verwendung dieses Kriteriums auf Studien, die zum einen auf vollständigen Datenmatrizen beruhen und zum anderen nur Imputationsverfahren vergleichen. Die meisten anderen Kriterien besitzen diese Einschränkungen nicht.

Als zweithäufigstes werden die Verfahren anhand ihrer Auswirkungen auf lineare Regressionsmodelle bewertet. Auch hierbei variiert je nach Studie die konkrete Ausgestaltung der Beurteilung. Ein Teil der Studien untersucht die Auswirkungen auf die Parameterschätzungen (und deren Standardfehler). Ein anderer Teil zieht zur Beurteilung die Auswirkungen auf die Modellgüte (u. a. anhand von geschätzten R^2-Werten oder durch den mittleren quadratischen Fehler der Prognose) heran.

Neben diesen beiden Kriterien wird eine Vielzahl anderer Gütemaße zur Beurteilung der Imputationsverfahren eingesetzt. Diese stammen häufig aus dem Bereich der Statistik, wie die Beurteilung anhand verschiedener Parameterschätzungen und statistischer Modelle, oder aus dem Bereich des maschinellen Lernens. Auch bei diesen Kriterien lässt sich vielmals feststellen, dass die Operationalisierung von Studie zu Studie variiert.

Zusätzlich zu diesen Unterschieden in den Details der Kriterien zeigt die Abbildung 5, dass sehr viele verschiedene Kriterien zur Bewertung von Imputationsverfahren verwendet werden. Es existiert jedoch kein Kriterium, welches allgemein akzeptiert ist und immer herangezogen wird. So wird selbst das am häufigsten verwendete Kriterium in nicht einmal einem Drittel der Studien verwendet. Hierdurch ergibt sich zunächst ein heterogenes Bild,

[6] Ein Teil der Parameterschätzverfahren führen als Zwischenschritt auch eine Imputation durch. Jedoch dienen die imputierten Werte nur der Parameterschätzung und werden häufig nicht ausgegeben (vgl. Little und Rubin 2002, S. 166–167).

gleichzeitig ermöglicht dies aber, die Imputationsverfahren aus unterschiedlichen Blickrichtungen zu betrachten.

2.5 Diskussion der Untersuchungsdesigns

Neben dem grundsätzlichen Unterschied im Vorgehen bei der Verwendung von unvollständigen und vollständigen Daten zeigen die vorherigen Kapitel, dass sich viele Studien zusätzlich in ihren Details unterscheiden. Dies erschwert den direkten Vergleich einzelner Simulationsstudien. Ein solcher Vergleich einzelner Studien soll im Rahmen dieser Arbeit auch nicht durchgeführt werden. Vielmehr ist das Ziel, Aussagen über die Güte von verschiedenen Imputationsverfahren abzuleiten. Hierfür bilden die unterschiedlichen Aspekte, die in den bestehenden Studien untersucht werden, eine breite Basis, bei der individuelle Unterschiede zwischen den Studien eher vorteilhaft als nachteilig sind.

Dennoch ist für diese Untersuchung wichtig, dass die einbezogenen Studien verlässliche Aussagen über die Güte der Imputationsverfahren liefern. Wie bereits im vorherigen Kapitel angesprochen, ist dies bei der Verwendung von unvollständigen Datenmatrizen meist nicht möglich, da hier das Verhalten der Verfahren nur deskriptiv untersucht werden kann.

Außerdem ist die Studienverlässlichkeit auch bei der Verwendung von vollständigen Datenmatrizen problematisch, wenn im Rahmen einer Simulation zu wenige Wiederholungen durchgeführt werden. Dieses Problem wird – wenn überhaupt – in der Literatur nur am Rande beachtet, ist jedoch bei genauerer Betrachtung von wesentlicher Bedeutung. So mussten unter anderem auf Grund einer zu geringen Anzahl an Wiederholungen schon publizierte Studienergebnisse korrigiert werden (vgl. Kim et al. 2005, Kim et al. 2006). Ferner sind veröffentliche Ergebnisse mit einer geringen Anzahl an Wiederholungen sehr instabil und zum Teil widersprüchlich (vgl. z. B. die Diagramme in Farhangfar et al. (2008, S. 3701–3703)).

In der Abbildung 6 ist die Anzahl an Wiederholungen für die untersuchten Studien mit vollständigen Datenmatrizen visualisiert.[7] Studien, bei denen die Anzahl an Wiederholungen weder im Text angegeben noch aus der Beschreibung ableitbar ist, sind in der Abbildung unter dem Punkt „NA“ aufgeführt. Aus der Abbildung 6 ist ersichtlich, dass als Wiederholungszahl häufig 1, 10, 100 oder 1000 gewählt wird. Hingegen sind mehr als 1000 Wiederholungen selten.

[7] Alle untersuchten Studien, die nur unvollständige Datenmatrizen verwenden, führen keine Wiederholungen durch. Daher werden diese Studien in der Abbildung nicht berücksichtigt.

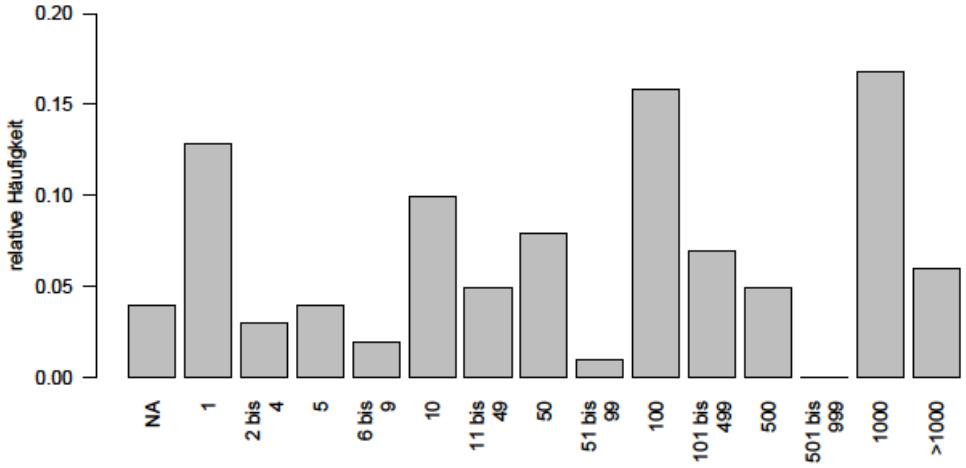

Abbildung 6: Anzahl der Wiederholungen bei der Verwendung vollständiger Datenmatrizen

Die Wahl der Anzahl an Wiederholungen wird in der Literatur häufig nicht begründet. Ausnahmen stellen Marshall et al. (2010a), Marshall et al. (2010b) und Laaksonen (2003) dar. So wählen Marshall et al. (2010a) 500 Wiederholungen, um eine Genauigkeit von mindestens 5 % bei sechs von acht Regressionskoeffizienten zu erhalten. In einer weiteren Studie entscheiden sich Marshall et al. (2010b) auf Grund ähnlicher Überlegungen für 1000 Wiederholungen. Ferner stellt Laaksonen (2003, S. 1014) fest, dass seine Ergebnisse sich nach 60 Wiederholungen nicht mehr stark ändern. Trotzdem verwendet Laaksonen (2003, S. 1014) 130 Simulationsläufe, um zusätzliche Sicherheit zu erhalten. An diesen drei Begründungen für die Auswahl der Wiederholungszahl zeigt sich, dass es keine allgemeingültige Mindestanzahl an benötigten Wiederholungen gibt. Jedoch deuten die dargelegten Sachverhalte daraufhin, dass bei einer zu geringen Anzahl an Wiederholungen das Risiko besteht, Ergebnisse zu verfälschen.

3 Analyse der MD-Verfahren anhand der untersuchten Studien

Im Folgenden werden die Ergebnisse der verschiedenen Studien aggregiert. In diese Aggregation fließen zusätzlich die MD-Verfahren ein, die keine Imputationsverfahren sind, damit die Ergebnisse der Studien möglichst vollständig widergespiegelt werden. Um ein unverfälschtes Bild der MD-Verfahren zu erhalten, werden Studien mit weniger als 100 Wiederholungen auf Grund der im vorherigen Kapitel beschriebenen Problematik von der

weiteren Betrachtung ausgeschlossen.[8] Dadurch verbleiben für die weitere Untersuchung 49 der 125 ursprünglich betrachteten Studien.

Jeder der 49 Studien gibt entweder eine Rangfolge der untersuchten MD-Verfahren an oder eine solche lässt sich aus den Ergebnissen ableiten.[9] Anhand dieser Rangfolgen werden die Verfahren untersucht. Im Folgenden werden zunächst die Ergebnisse der einzelnen Verfahren getrennt betrachtet. Anschließend werden im Kapitel 3.2 die MD-Verfahren paarweise anhand der Studien verglichen. Abschließend wird noch der Einfluss, den verschiedene Faktoren auf die MD-Verfahren haben, untersucht.

3.1 Ergebnisse der einzelnen MD-Verfahren

In diesem Kapitel werden die Resultate jedes Verfahrens einzeln analysiert. Dazu werden alle Studien herangezogen, in denen das betrachtete Verfahren untersucht wird. Für jede Studie wird der Rang des Verfahrens im Vergleich zu allen anderen einbezogenen Verfahren bestimmt. So wird ein Überblick über die Ergebnisse der einzelnen Verfahren geschaffen, ohne dabei auf die Vergleichsverfahren im Detail einzugehen, da diese von Studie zu Studie variieren. Jedoch wird bei der Untersuchung berücksichtigt, wie viele Verfahren eine Studie vergleicht, da z. B. die Interpretation eines zweiten Ranges unter anderem abhängig von der Anzahl der Vergleichsverfahren ist.

Die Ergebnisse jedes Verfahrens werden in Form eines Sterndiagramms visualisiert. Diese Diagramme befinden sich in der Abbildung 7. In der Abbildung wird jedes Verfahren gezeigt, das in mindestens drei der 49 Studien untersucht wird. Für jedes Verfahren werden drei Polygone dargestellt. Dabei zeigt das rote Polygon die Ränge, die ein Verfahren im Rahmen der Simulationsstudie erreicht. Das blaue Polygon gibt an, wie viele unterschiedliche Verfahren in den Vergleich eingeflossen sind.[10] Das schwarze Polygon dient der „Normierung“. Die genaue Konstruktion bzw. Interpretation der Sterndiagramme geschieht folgendermaßen:

---

8 Die Grenze von mindestens 100 Wiederholungen ist ein Kompromiss zwischen der Sicherheit, dass eine Studie verlässliche Ergebnisse liefert, und dem nicht übermäßigen Ausschluss von Studien. Sie orientiert sich an den Ergebnissen von Laaksonen (2003, S. 1014) .

9 Falls ein Verfahren in einer Studie mit verschiedenen Parametereinstellungen untersucht wird, wird in der Rangfolge nur das Ergebnis der besten Parametereinstellung berücksichtigt. Hierdurch wird das Potential der Verfahren in den Vordergrund gerückt, welches durch eine optimale Parameterwahl erreichbar ist.

10 Falls ein Verfahren in einer Studie mit verschiedenen Parametereinstellungen untersucht wird, wird es bei der Anzahl an unterschiedlichen Verfahren nur einmal berücksichtigt, da ein solches Verfahren auch nur einmal in der Rangfolge erfasst wird.

- Jede Studie, in der ein Verfahren untersucht wird, wird durch eine Ecke in jedem der drei farbigen Polygone repräsentiert. Falls beispielsweise ein Verfahren in acht Studien untersucht wird, sind alle drei Polygone Achtecke. Das heißt, je mehr Ecken die Polygone besitzen, (bzw. je „runder“ das schwarze Polygon ist), desto häufiger wird das Verfahren untersucht.
- Das schwarze Polygon dient der „Normierung“. Jede Ecke des Polygons befindet sich auf der maximalen Anzahl an unterschiedlichen Verfahren über alle Studien. Hierdurch werden die Gesamtzahl an Verfahren in einer Studie und der Rang des betrachteten Verfahrens in Bezug zur maximal untersuchten Anzahl an Verfahren über alle Studien gesetzt.
- Die Ecken des blauen Polygons geben die Anzahl unterschiedlicher Verfahren in einer Studie an. Falls beispielsweise fünf Verfahren in einer Studie untersucht werden, dann befindet sich der blaue Punkt auf der Koordinate fünf für diese Studie. Je weiter „außen“ (bezogen auf den „Mittelpunkt“ des Sterndiagramms) ein Punkt des blauen Polygons liegt, desto mehr Verfahren werden in der zugehörigen Studie untersucht.
- Jede Ecke des roten Polygons steht für den Rang des jeweiligen Verfahrens in einer Studie. Je besser ein Verfahren abschneidet, desto geringer ist der Rang in der Studie und desto weiter „innen“ (bezogen auf das blaue bzw. schwarze Polygon) liegen die Punkte des roten Polygons. Je kleiner also der Flächeninhalt des roten Polygons ist, desto besser schneidet das Verfahren insgesamt ab.
- Das Verhältnis des Flächeninhalts von dem roten Polygon zu dem des blauen ist ein Maß für den relativen Rang eines Verfahrens. Je kleiner der Flächeninhalt des roten Polygons im Vergleich zu dem des blauen ist, desto mehr Verfahren sind in den Vergleichen schlechter als das durch das rote Polygon dargestellte Verfahren. Ein gutes Verfahren zeichnet sich also nicht nur durch einen kleinen Flächeninhalt des roten, sondern auch durch einen gleichzeitig großen Flächeninhalt des blauen Polygons aus.

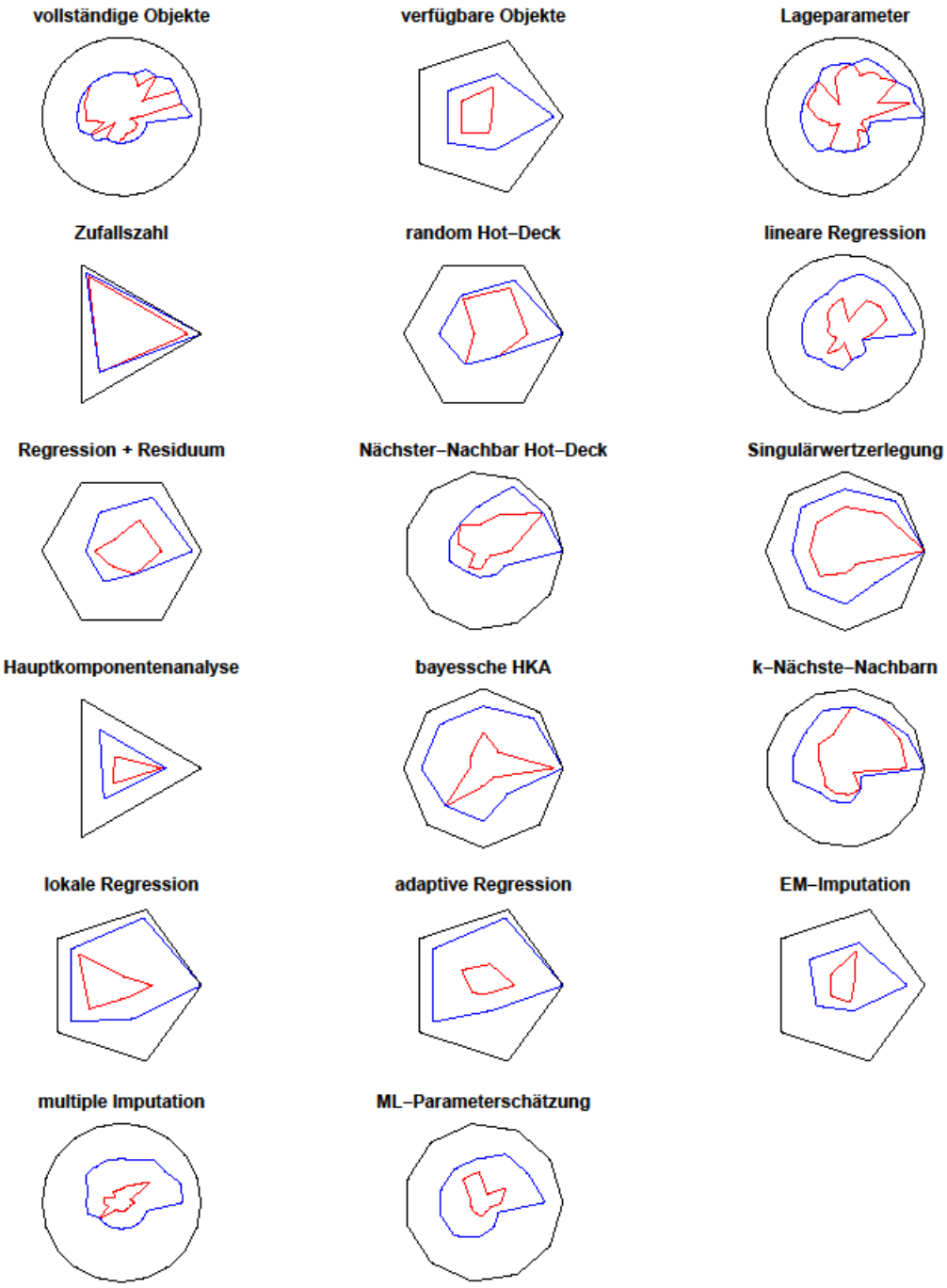

Abbildung 7: Ergebnisse der MD-Verfahren

Die Verfahren in Abbildung 7 sind wie folgt geordnet: Eliminierungsverfahren, Imputationsverfahren, multiple Imputation und ML-Parameterschätzverfahren. Die ersten beiden Verfahren, Analyse der vollständigen Objekte und Analyse der verfügbaren Objekte, zählen zu den Eliminierungsverfahren. Aus der Abbildung 7 wird deutlich, dass die Elimination aller unvollständigen Objekte meist eine der schlechtesten Optionen ist. Jedoch existieren auch vereinzelte Studien, in denen dieses Vorgehen gute Ergebnisse liefert. Das andere Eliminierungsverfahren, die Analyse der verfügbaren Objekte, wird nur in fünf Studien un-

tersucht. Dabei erzielt es in einer Studie sehr gute Ergebnisse und in den restlichen vier liegt es im Mittelfeld.

Die einfachen Imputationsverfahren (Lageparameterimputation, Zufallszahlimputation und random Hot-Deck) sind nicht in der Lage, konstant gute Ergebnisse zu liefern. So gehört die Imputation eines Lageparameters zwar in einzelnen Studien zu den besten Verfahren, jedoch führt sie in vielen Studien zu sehr schlechten Ergebnissen. Besonders deutlich unterlegen ist die Imputation einer Zufallszahl, die in keiner der drei Studien zu akzeptablen Ergebnissen führt. Auch das random Hot-Deck gehört in den untersuchten Studien meist zum Mittelfeld oder zu den Verfahren mit schlechten Ergebnissen.

Ein besseres Bild bieten die drei klassischen Imputationsverfahren lineare Regression, lineare Regression plus Residuum und Nächster-Nachbar Hot-Deck. Insbesondere die Imputation mittels linearer Regression befindet sich häufig im Mittelfeld oder liefert gute Ergebnisse. Die zusätzliche Addition eines Residuums führt zu etwas schlechteren Ergebnissen, ist aber häufig auch im Mittelfeld zu finden. Hingegen ist das Bild beim Nächster-Nachbar Hot-Deck nicht eindeutig. Das Verfahren führt teils zu guten, teils zu schlechten Ergebnissen.

Die nächsten drei Imputationsverfahren Singulärwertzerlegung, Hauptkomponentenanalyse und bayessche Hauptkomponentenanalyse basieren auf Repräsentationsverfahren und erzielen sehr unterschiedliche Ergebnisse. Die Imputation mittels Singulärwertzerlegung gehört oftmals zum schlechten Mittelfeld in den untersuchten Studien. Bei der Imputation mittels Hauptkomponentenanalyse sind die Ergebnisse der drei Studien widersprüchlich. In einer Studie gehört dieses Verfahren zu den besten, in einer anderen zu den schlechtesten und in der dritten Studie liegt es im Mittelfeld. Ein ähnliches Bild ergibt sich bei der Imputation mittels bayesscher Hauptkomponentenanalyse, welche teilweise gute und teilweise schlechte Ergebnisse erzielt.

Auch die Ergebnisse der drei lokalen Verfahren k-Nächste-Nachbarn, lokale Regression und adaptive Regression[11] sind sehr unterschiedlich. So führt die Imputation mittels k-Nächste-Nachbarn zu mittelmäßigen oder schlechten Ergebnissen, während die lokale Re-

[11] Das hier aufgeführte Verfahren der adaptiven Regression basiert auf einem Vorschlag von Bø et al. (2004). Bø et al. (2004) bezeichnen dieses Verfahren in ihrer ursprünglichen Veröffentlichung als LSimpute_adaptive. Bei anderen Autoren ist es unter dem Begriff least squares adaptive (LSA) zu finden (vgl. z. B. Brock et al. 2008, Chiu et al. 2013).

gression meist im Mittelfeld liegt. Im Gegensatz dazu gehört die adaptive Regression in den Studien, in denen sie mit einbezogen wird, zu den besten Verfahren.

Die letzten drei Verfahren EM-Imputation, multiple Imputation und Verfahren, die auf ML-Parameterschätzungen[12] basieren, erzielen in der Regel gute Ergebnisse. Dabei sind insbesondere die Ergebnisse der multiplen Imputation hervorzuheben, die fast immer zu den besten zählen. Aber auch die Ergebnisse der ML-basierten Verfahren und der EM-Imputation sind häufig gut oder liegen zumindest im Mittelfeld.

MD-Verfahren	Studien	$\frac{\text{rot}}{\text{blau}}$	$\frac{\text{ØRang}}{\text{ØVerfahren}}$
vollständige Objekte	29	0.69 (16)	0.79 (16)
verfügbare Objekte	5	0.22 (5)	0.46 (5)
Lageparameter	31	0.58 (14)	0.73 (14)
Zufallszahl	3	0.84 (17)	0.91 (17)
random Hot-Deck	6	0.48 (13)	0.71 (13)
lineare Regression	19	0.34 (10)	0.56 (8)
Regression + Residuum	6	0.33 (9)	0.58 (10)
Nächster-Nachbar Hot-Deck	11	0.39 (11)	0.61 (11)
Singulärwertzerlegung	8	0.47 (12)	0.69 (12)
Hauptkomponentenanalyse	3	0.29 (7)	0.57 (9)
bayessche HKA	8	0.30 (8)	0.51 (7)
k-Nächste-Nachbarn	13	0.61 (15)	0.75 (15)
lokale Regression	5	0.24 (6)	0.47 (6)
adaptive Regression	5	0.12 (1)	0.34 (1)
EM-Imputation	5	0.19 (4)	0.46 (4)
multiple Imputation	20	0.15 (2)	0.37 (2)
ML-Parameterschätzung	11	0.17 (3)	0.40 (3)

Tabelle 3: Ergebnisse der einzelnen MD-Verfahren

Die Ergebnisse der Abbildung 7 werden in der Tabelle 3 zusammengefasst. Die Tabelle gibt für jedes MD-Verfahren die Anzahl an Studien, in die es mit einbezogen wird und das Verhältnis des Flächeninhalts vom roten zum blauen Polygon an. Im Klammern hinter diesem Verhältnis steht der Rang des Verfahrens, falls die Güte der Verfahren nur anhand dieser Kennzahl bewertet wird.

Eine andere Möglichkeit die Verfahren anhand einer Kennzahl zu bewerten, ist die Betrachtung des durchschnittlichen Rangs über alle Studien, in denen ein Verfahren untersucht wird. Jedoch besitzt diese Kennzahl eine ähnliche Problematik wie die alleinige Be-

[12] Hierunter zählen vor allem der EM-Algorithmus und auf Full Information Maximum Likelihood (FIML) basierende Verfahren.

trachtung des roten Polygons. Sie erfasst nicht, wie viele Verfahren in die Vergleiche einfließen. Um dies zu berücksichtigen wird der durchschnittliche Rang durch die durchschnittliche Anzahl an unterschiedlichen Verfahren, die in den Studien des betrachteten Verfahrens verwendet werden, dividiert. Diese Kennzahl ist zusammen mit der Position, die ein Verfahren bei Betrachtung nur dieser Kennzahl erreicht, in der Tabelle 3 angegeben.

Die Ergebnisse in der Tabelle 3 unterstützen die zur Abbildung 7 getroffenen Aussagen. So erzielen die Verfahren, die bereits anhand der Abbildung 7 als gut eingestuft wurden, sowohl niedrige Ergebnisse bei dem Verhältnis der Flächeninhalte als auch bei dem Verhältnis aus Rang und Verfahrensanzahl. Die Reihenfolge der besten Verfahren ist bei beiden nahezu identisch. Insbesondere führen beide Kennzahlen zu derselben Reihenfolge bei den sechs besten und sechs schlechtesten Verfahren.

Insgesamt zeigen die Abbildung 7 und die Tabelle 3, dass die adaptive Regression, die multiple Imputation und die ML-Parameterschätzverfahren in den Vergleichen, in die sie mit einbezogen werden, immer zu den besten Verfahren zählen. Auf der anderen Seite ist die Zufallszahlimputation das mit Abstand schlechteste Verfahren. Aber auch die Analyse der vollständigen Objekte, die k-Nächste-Nachbarn-Imputation und die Lageparameterimputation zählen in den untersuchten Studien überwiegend zu den schlechtesten Verfahren. Die Ergebnisse der restlichen Verfahren bewegen sich normalerweise zwischen diesen beiden Gruppen.

3.2 Paarvergleich der MD-Verfahren

Das vorherige Kapitel gibt bereits einen Überblick über die Ergebnisse der Verfahren in den Vergleichsstudien. Aber weder die Abbildung 7 noch die Tabelle 3 erfasst, welche Verfahren in den Studien miteinander verglichen werden. Die Auswahl der Vergleichsverfahren beeinflusst jedoch die Platzierung eines Verfahrens. Beispielsweise wird ein mittelmäßiges Verfahren, welches nur mit schlechten Verfahren verglichen wird, besser erscheinen als ein ähnliches Verfahren, welches nur mit guten Verfahren verglichen wird. Aus diesem Grund werden in diesem Kapitel die Verfahren paarweise miteinander verglichen.

Dazu wird zunächst für jedes Verfahren erfasst, wie häufig es mit allen anderen Verfahren verglichen wird. Ferner wird festgehalten, in wie vielen Studien ein Verfahren mindestens genauso gut ist wie ein Vergleichsverfahren. Hieraus wird schließlich der Anteil an Stu-

dien berechnet, in denen ein Verfahren genauso gut oder besser als ein Vergleichsverfahren abschneidet. Das Ergebnis ist in der Tabelle 4 dargestellt. Die Prozentzahlen geben an, wie häufig das Verfahren in einer Zeile mindestens genauso gut ist wie das Verfahren in der zugehörigen Spalte. Falls zwei Verfahren in keiner der Studien gleichzeitig untersucht werden, steht an der entsprechenden Stelle ein „-“ in der Tabelle.

Mit Hilfe der Tabelle 4 lassen sich zwei Verfahren direkt miteinander vergleichen (sofern ein solcher Vergleich in mindestens einer Studie durchgeführt wird). Außerdem kann ein Verfahren sowohl anhand der zum Verfahren gehörenden Zeile als auch anhand der zugehörigen Spalte beurteilt werden. Je höher die Prozentzahlen in der Zeile eines Verfahren sind, desto häufiger schneidet es mindestens so gut wie die Vergleichsverfahren ab. Umgekehrt ist ein Verfahren umso besser, je kleiner die Prozentzahlen in der zugehörigen Spalte sind. Diese Interpretation der Spaltenwerte basiert auf folgendem Zusammenhang: In der Zelle in Zeile A und Spalte B steht, wie häufig das Verfahren A mindestens genauso gut wie das Verfahren B ist. Also gibt die Differenz aus Eins und dem Zellenwert an, wie häufig das Verfahren A schlechter als das Verfahren B ist, oder anders formuliert: Wie häufig das Verfahren B besser als das Verfahren A ist. Je kleiner die Einträge in der zu B gehörigen Spalte sind, desto häufiger ist das Verfahren B anderen Verfahren überlegen.

Beispielsweise ist das Verfahren A mindestens genauso gut wie das Verfahren B, wenn die Zelle in Zeile A und Spalte B eine Prozentzahl größer gleich 50 % aufweist, da in diesem Fall das Verfahren A in wenigstens der Hälfte der Vergleiche dem Verfahren B mindestens ebenbürtig ist. Falls zusätzlich in der Zelle mit Zeile B und Spalte A eine Prozentzahl kleiner als 50 % steht, ist das Verfahren A dem Verfahren B überlegen, da das Verfahren B in mehr als der Hälfte der Studien zu schlechteren Ergebnissen als das Verfahren A führt. Falls in einer Studie die Verfahren A und B vergleichbare Ergebnisse erzielen, erhöht dies sowohl die Prozentzahl in der Zeile von A als auch in der Zeile von B in den entsprechenden Spalten. Daher ist in diesem Fall die Summe von der Zelle in Zeile A, Spalte B und der Zelle in Zeile B, Spalte A größer als 100 %. Aus demselben Grund steht auf der Hauptdiagonalen der Tabelle 4, auf der jedes Verfahren mit sich selbst verglichen wird, immer 100 %.

Auf je mehr Studien die Paarvergleiche von zwei Verfahren in der Tabelle 4 basieren, desto verlässlicher sind sie. Diese Anzahl an Vergleichen ist zu jedem Vergleich in Klammern angegeben. Ferner ist auf der Hauptdiagonalen in Klammern angeführt, in wie vielen Studien ein Verfahren insgesamt untersucht wird.

	CC	AC	LP	ZZ	RHD	LR	RR	NNHD	SWZ	HKA	BHKA	kNN	LoR	AR	EMI	MI	ML
CC	100%	50%	37%	100%	67%	43%	75%	40%	0%	0%	0%	0%	-	-	0%	7%	0%
	(29)	(4)	(19)	(2)	(3)	(14)	(4)	(5)	(1)	(2)	(1)	(1)			(1)	(15)	(8)
AC	50%	100%	80%	100%	100%	100%	100%	100%	-	0%	-	-	-	-	100%	0%	0%
	(4)	(5)	(5)	(1)	(1)	(3)	(1)	(1)		(1)					(1)	(1)	(1)
LP	68%	20%	100%	100%	50%	38%	0%	57%	20%	100%	0%	25%	33%	0%	25%	0%	33%
	(19)	(5)	(31)	(3)	(4)	(13)	(4)	(7)	(5)	(3)	(5)	(8)	(3)	(2)	(4)	(11)	(6)
ZZ	0%	0%	33%	100%	0%	0%	0%	0%	-	-	-	-	-	-	-	-	-
	(2)	(1)	(3)	(3)	(1)	(2)	(1)	(2)									
RHD	33%	0%	75%	100%	100%	33%	0%	0%	-	-	-	-	-	-	-	0%	0%
	(3)	(1)	(4)	(1)	(6)	(3)	(2)	(3)								(1)	(1)
LR	57%	33%	69%	100%	67%	100%	80%	75%	100%	0%	0%	100%	100%	0%	0%	20%	20%
	(14)	(3)	(13)	(2)	(3)	(19)	(5)	(4)	(1)	(1)	(2)	(2)	(2)	(2)	(1)	(5)	(5)
RR	25%	0%	100%	100%	100%	40%	100%	67%	-	-	-	-	-	-	-	0%	50%
	(4)	(1)	(4)	(1)	(2)	(5)	(6)	(3)								(2)	(2)
NNHD	80%	0%	57%	100%	100%	25%	67%	100%	0%	-	0%	0%	0%	0%	-	0%	0%
	(5)	(1)	(7)	(2)	(3)	(4)	(3)	(11)	(1)		(1)	(1)	(1)	(1)		(3)	(1)
SWZ	100%	-	80%	-	-	100%	-	100%	100%	100%	17%	57%	33%	0%	100%	100%	-
	(1)		(5)			(1)		(1)	(8)	(1)	(6)	(7)	(3)	(3)	(1)	(1)	
HKA	100%	100%	100%	-	-	100%	-	-	0%	100%	-	-	-	-	50%	-	-
	(2)	(1)	(3)			(1)			(1)	(3)					(2)		
BHKA	100%	-	100%	-	-	100%	-	100%	83%	-	100%	75%	60%	50%	0%	100%	-
	(1)		(5)			(2)		(1)	(6)		(8)	(8)	(5)	(4)	(1)	(1)	
kNN	100%	-	75%	-	-	0%	-	100%	71%	-	25%	100%	0%	0%	0%	33%	-
	(1)		(8)			(2)		(1)	(7)		(8)	(13)	(5)	(5)	(2)	(3)	
LoR	-	-	67%	-	-	50%	-	100%	100%	-	60%	100%	100%	50%	0%	-	-
			(3)			(2)		(1)	(3)		(5)	(5)	(5)	(4)	(1)		
AR	-	-	100%	-	-	100%	-	100%	100%	-	100%	100%	75%	100%	50%	-	-
			(2)			(2)		(1)	(3)		(4)	(5)	(4)	(5)	(2)		
EMI	100%	0%	100%	-	-	100%	-	-	100%	100%	100%	100%	100%	50%	100%	0%	0%
	(1)	(1)	(4)			(1)			(1)	(2)	(1)	(2)	(1)	(2)	(5)	(1)	(1)
MI	93%	100%	100%	-	100%	100%	100%	100%	100%	-	0%	100%	-	-	100%	100%	86%
	(15)	(1)	(11)		(1)	(5)	(2)	(3)	(1)		(1)	(3)			(1)	(20)	(7)
ML	100%	100%	83%	-	100%	80%	50%	100%	-	-	-	-	-	-	100%	71%	100%
	(8)	(1)	(6)		(1)	(5)	(2)	(1)							(1)	(7)	(11)

In der Tabelle ist angegeben, wie häufig das Verfahren in einer Zeile mindestens genauso gut ist wie das Verfahren in einer Spalte. In Klammern ist die Anzahl an Studien aufgeführt, in denen die Verfahren verglichen werden. Die Abkürzungen in der Tabelle bedeuten: Analyse der vollständigen Objekte (Complete Case Analysis) (CC), Analyse der verfügbaren Objekte (Available Case Analysis) (AC), Lageparameterimputation (LP), Zufallszahlimputation (ZZ), random Hot-Deck (RHD), lineare Regression (LR), Regression plus Residuum (RR), Nächster-Nachbar Hot-Deck (NNHD), Singulärwertzerlegung (SWZ), Hauptkomponentenanalyse (HKA), bayessche Hauptkomponentenanalyse (BHKA), k-Nächste-Nachbarn (kNN), lokale Regression (LoR), adaptive Regression (AR), EM-Imputation (EMI), multiple Imputation (MI) und auf ML-Parameterschätzverfahren basierende Verfahren (ML).

Tabelle 4: Vergleich der MD-Verfahren

Die Reihenfolge der Verfahren in der Tabelle 4 entspricht der Reihenfolge in der Abbildung 7. Auf die Verfahren wird zunächst in dieser Reihenfolge eingegangen. Das erste Verfahren, die Analyse der vollständigen Objekte, kann sich gegenüber der Zufallszahlimputation, Imputation mittels Regression plus Residuum und – weniger deutlich – gegenüber dem random Hot-Deck behaupten. Die restlichen elf Verfahren, mit denen die Analyse der vollständigen Objekte verglichen wird, sind entweder genauso gut wie diese oder besser. Das zweite Eliminierungsverfahren, die Analyse der verfügbaren Objekte, ist sieben Verfahren im direkten Vergleich überlegen und vier Vergleichsverfahren sind besser oder genauso gut wie sie. Jedoch basieren alle Vergleiche, mit Ausnahme des Vergleichs mit der Lageparameterimputation und der Analyse der vollständigen Objekte, auf drei oder weniger Studien.

Die drei einfachen Imputationsverfahren Lageparameterimputation, Zufallszahlimputation und random Hot-Deck können sich größtenteils nicht gegenüber anderen MD-Verfahren durchsetzen. So ist die Zufallszahlimputation allen Verfahren unterlegen, mit denen sie verglichen wird. Außerdem kann das random Hot-Deck sich nur gegenüber der Zufallszahlimputation und bedingt gegenüber der Lageparameterimputation durchsetzen. In allen anderen Vergleichen ist die random Hot-Deck-Imputation den anderen Verfahren unterlegen. Ferner erzielt die Lageparameterimputation, mit Ausnahme der Zufallszahlimputation und der Imputation mittels Hauptkomponentenanalyse, keine besseren Ergebnisse als die Vergleichsverfahren. Zwölf der sechzehn Paarvergleich mit der Lageparameterimputation basieren auf vier oder mehr Studien, drei Paarvergleiche basieren auf drei und ein Paarvergleich auf zwei Studien. Hingegen werden alle Vergleiche der anderen beiden Verfahren höchstens dreimal untersucht. Die einzige Ausnahme hiervon ist der Vergleich zwischen dem random Hot-Deck und der Lageparameterimputation, welcher auf vier Studien basiert.

Die drei Imputationsverfahren lineare Regression, Regression plus Residuum und Nächster-Nachbar Hot-Deck sind besser als die drei einfachen Imputationsverfahren. Ferner ist die Imputation mittels linearer Regression nicht nur den beiden anderen Verfahren überlegen, sondern auch der Singulärwertzerlegung, k-Nächsten-Nachbarn-Imputation und der Imputation mittels lokaler Regression. Jedoch ist die lineare Regression sieben anderen Verfahren unterlegen. Die Imputation mittels Regression plus Residuum ist lediglich den drei einfachen Imputationsverfahren und dem Nächsten-Nachbar Hot-Deck mindestens ebenbürtig. Auch das Nächster-Nachbar Hot-Deck kann sich zusätzlich zu den einfachen Imputationsverfahren nur noch gegenüber der Analyse der vollständigen Objekte durchset-

zen. Erneut ist zu den Aussagen in diesem Absatz anzumerken, dass 27 der Vergleiche auf drei oder weniger Studien beruhen.

Die folgenden drei Verfahren Imputation mittels Singulärwertzerlegung, Hauptkomponentenanalyse und bayesscher Hauptkomponentenanalyse sind allen vorherigen Verfahren überlegen, sofern sie mit diesen verglichen werden. Die Imputation mittels Singulärwertzerlegung ist außerdem in den untersuchten Studien der Imputation mittels Hauptkomponentenanalyse, k-Nächste-Nachbarn sowie der EM-Imputation und der multiplen Imputation mindestens ebenbürtig. Jedoch ist sie der Imputation mittels bayesscher Hauptkomponentenanalyse, lokaler und adaptiver Regression unterlegen. Über 70 % der Vergleiche der Imputation mittels Singulärwertzerlegung basieren auf drei oder weniger Studien. Die Imputation mittels Hauptkomponentenanalyse wird nur in drei Studien überhaupt untersucht. Daher wird sie bisher nur mit wenigen Verfahren verglichen und Aussagen über diese Methode sind schwierig. Die Imputation mittels Hauptkomponentenanalyse ist jedoch zumindest der Lageparameterimputation in drei Studien ebenbürtig und der Analyse der vollständigen Objekte in zwei Studien überlegen. Die Datenlage für die bayesianische Variante ist etwas besser. Sie führt meist zu guten Ergebnissen und ist in den Paarvergleichen einzig der EM-Imputation unterlegen.

Die Ergebnisse der drei lokalen Ansätze k-Nächste-Nachbarn, lokale Regression und adaptive Regression sind sehr unterschiedlich. Die k-Nächste-Nachbarn-Imputation ist den anderen beiden lokalen Ansätzen und auch der linearen Regression, der EM-Imputation sowie der multiplen Imputation unterlegen. Demgegenüber ist die Imputation mittels lokaler Regression fast allen Verfahren, mit denen sie verglichen wird, mindestens ebenbürtig. Die Vergleichsresultate der adaptiven Regression übertreffen die der lokalen Regression nochmals. Die Imputation mittels adaptiver Regression ist mindestens genauso gut wie jedes Verfahren, mit dem sie verglichen wird, und fast immer sogar besser. Es existieren allerdings keine direkten Vergleiche der Imputation mittels adaptiver Regression mit acht anderen Verfahren. Auch für die Imputation mittels lokaler Regression fehlen Vergleiche mit acht der anderen Verfahren. Insbesondere existieren keine Vergleiche zwischen diesen beiden und den ML-basierten Verfahren sowie der multiplen Imputation.

Auch die EM-Imputation, multiple Imputation und ML-basierten Verfahren führen größtenteils zu sehr guten Ergebnissen. So muss sich die EM-Imputation in den untersuchten Studien nur der multiplen Imputation, den ML-basierten Verfahren und der Analyse der verfügbaren Objekten geschlagen geben, wobei diese Ergebnisse jeweils nur durch eine

Studie gestützt werden. Noch besser sind die Resultate der multiplen Imputation. Sie ist allen Verfahren mindestens ebenbürtig (mit Ausnahme der Imputation mittels bayesscher Hauptkomponentenanalyse, wobei hierfür lediglich eine Studie existiert). Außerdem fallen auch die Ergebnisse der ML-basierten Verfahren gut aus.

Wie bei den Ergebnissen der Verfahren bereits angemerkt, basieren viele Vergleiche in der Tabelle 4 auf nur wenigen Studien oder zwei Verfahren werden in keiner der Studien verglichen. Dieses lückenhafte Bild wird besonders deutlich daran, dass 45 von den 136 möglichen Paarvergleichen in keiner der Studien untersucht werden. Ferner basieren 59 der 90 untersuchten Paarvergleiche auf höchstens drei Studien. Für über 75 % der Vergleiche existieren folglich überhaupt keine oder nur wenige Studien.

Diese schwache Datenbasis ist für die ersten fünf einfachen Verfahren in der Tabelle nicht weiter problematisch. Hier weisen die bestehenden Ergebnisse deutlich darauf hin, dass die fehlenden Vergleiche zu Ungunsten dieser sechs Verfahren ausgehen würden. Für andere Verfahren, wie z. B. die multiple Imputation und die Imputation mittels adaptiver Regression, ist es jedoch ohne weitere Studien schwer abzuschätzen, wie sie sich im direkten Vergleich verhalten.

3.3 Zusammenfassende Bewertung der MD-Verfahren

Aufbauend auf den beiden vorherigen Abschnitten werden in diesem Abschnitt die wichtigsten Erkenntnisse der untersuchten Studien zusammengefasst. Die drei besten Verfahren sowohl beim Paarvergleich als auch bei der Einzelbetrachtung sind die multiple Imputation, die Imputation mittels adaptiver Regression und die auf ML-Theorie basierenden Verfahren. Hierbei deuten die Ergebnisse in der Tabelle 4 daraufhin, dass die multiple Imputation den ML-basierten Verfahren leicht überlegen ist. Da in keiner der untersuchten Studien ein direkter Vergleich zwischen der Imputation mittels adaptiver Regression und der multiplen Imputation bzw. ML-basierten Verfahren durchgeführt wird, sind ähnliche Aussagen für die Imputation mittels adaptiver Regression nicht ableitbar.

Basierend auf den Ergebnissen der beiden vorherigen Abschnitte ist das schlechteste Verfahren die Zufallszahlimputation. Bei den Paarvergleichen zeigt sich außerdem, dass weitere einfache Verfahren, wie die Lageparameterimputation, das random Hot-Deck und die Analyse der vollständigen Objekte den meisten anderen Verfahren unterlegen sind. Jedoch deuten die Ergebnisse in der Abbildung 7 darauf hin, dass es durchaus Situationen gibt, in denen diese drei einfachen Verfahren zu akzeptablen Resultaten führen.

Die Ergebnisse der restlichen Verfahren bewegen sich überwiegend zwischen den beiden obigen Gruppen. In der Tabelle 4 zeigen sich die EM-Imputation und die Imputation mittels bayesscher Hauptkomponentenanalyse noch als zwei weitere bessere Verfahren. Dieses Ergebnis wird bei der EM-Imputation zusätzlich durch die Abbildung 7 gestützt. Demgegenüber zeigt sich bei der Imputation mittels bayesscher Hauptkomponentenanalyse in der Abbildung 7 ein widersprüchliches Bild. So schneidet sie teils gut, teils schlecht ab.

Allgemein zeigen die Auswertungen in der Abbildung 7 und der Tabelle 4, dass die Ergebnisse von vielen Verfahren nicht eindeutig sind. Ferner wird insbesondere aus der Tabelle 4 deutlich, dass viele Verfahren noch nie oder nur sehr selten miteinander verglichen wurden, was einen direkten Vergleich dieser Verfahren nahezu unmöglich macht. Es besteht hier noch erheblicher Forschungsbedarf.

3.4 Auswirkungen verschiedener Faktoren auf die Güte der MD-Verfahren

Neben dem Vergleich der MD-Verfahren liefern einige Simulationsstudien auch Erkenntnisse über den Einfluss verschiedener Faktoren auf die Güte der MD-Verfahren. In der Tabelle 5 ist aufgeschlüsselt, welchen Einfluss eine Erhöhung bzw. Verstärkung verschiedener Faktoren auf die in der jeweiligen Studie untersuchten MD-Verfahren hat.[13]

Falls ein Faktor in einer Studie nicht untersucht oder nicht systematisch variiert wird, ist die Studie der Spalte „nicht untersucht“ zugeordnet. Aus der Tabelle ist ersichtlich, dass nur der Einfluss der Ausfallrate in über der Hälfte der Studien systematisch untersucht wird. Alle anderen Faktoren werden in weniger als 35 % der Studien systematisch variiert. Die folgenden Aussagen beziehen sich jeweils auf die Studien, die den angesprochenen Faktor untersuchen.

	keinen	besser	schlechter	unterschiedlich	nicht untersucht
Anzahl Objekte	1	7	0	4	37
Anzahl Merkmale	0	2	1	2	44
Zusammenhang	1	3	1	7	37
Ausfallmechanismus[14]	0	0	8	9	32
Ausfallrate	1	1	23	3	21

Tabelle 5: Einfluss einer Erhöhung/Verstärkung eines Faktors auf die MD-Verfahren

[13] Es bestehen Wechselwirkungen zwischen einigen Faktoren und den MD-Verfahren. Dies kann dazu führen, dass die Wirkung eines einzelnen Faktors nicht eindeutig ist bzw. sich je nach MD-Verfahren anders verhält. Falls dies in einer Studie der Fall ist, wird die Studie in der Spalte „unterschiedlich“ aufgeführt.

[14] Hierbei wird ein MCAR-Mechanismus als die schwächste Form des Ausfalls und NMAR als stärkster Ausfallmechanismus angesehen.

Die Tabelle 5 zeigt, dass in sieben von zwölf Studien eine Erhöhung der Objektanzahl die Ergebnisse der MD-Verfahren generell verbessert. In vier Studien führt eine Erhöhung der Objektanzahl in einigen Fällen zu einer Verbesserung, aber in anderen Fällen auch zu einer Verschlechterung der Ergebnisse und in einer Studie hat die Objektanzahl keinen nennenswerten Einfluss auf die Ergebnisse. Die Auswirkungen der Anzahl an Merkmalen werden nur in fünf der 49 Studien systematisch untersucht und die Wirkungen dieses Faktors sind je nach Studie unterschiedlich. Auch die Auswirkungen einer Verstärkung des Zusammenhangs zwischen den Merkmalen sind nicht eindeutig.

Im Gegensatz dazu führt ein stärkerer Ausfallmechanismus in 8 von 17 Studien zu einer Verschlechterung der Ergebnisse und in 9 von 17 Studien ist die Auswirkung dieses Faktors nicht eindeutig. Aus der Tabelle 5 geht hervor, dass eine höhere Ausfallrate in über 80 % der Fälle die Ergebnisse der MD-Verfahren verschlechtert.

Zusammenfassend ergibt sich, dass eine höhere Anzahl an Objekten tendenziell zu besseren Ergebnissen der MD-Verfahren führt. Hingegen verschlechtert eine Verstärkung des Ausfalls, sowohl bezüglich des Anteils fehlender Werte als auch des Ausfallmechanismus, die Ergebnisse der MD-Verfahren. Bei den beiden Faktoren Anzahl der Merkmale und Zusammenhang zwischen den Merkmalen ergibt sich kein eindeutiges Bild. Dies liegt nicht zuletzt auch an der geringen Anzahl an Studien, die diese beiden Faktoren systematisch variieren.

4 Fazit

Das Ziel der vorliegenden Analyse war, die bereits existierenden Erkenntnisse zur Wahl eines geeigneten Imputationsverfahren zu aggregieren. Dabei zeigte sich, dass die drei Verfahren Imputation mittels adaptiver Regression, multiple Imputation und auf der ML-Theorie basierende Verfahren in den untersuchten Studien die besten Ergebnisse liefern. Jedoch existiert noch keine Studie, die die adaptive Regression direkt mit der multiplen Imputation oder mit auf ML-basierenden Verfahren vergleicht. Eine Beurteilung, welches der drei Verfahren am besten zur Behandlung fehlender Daten geeignet ist, ist daher nicht möglich. Dieses Problem der fehlenden oder nicht ausreichenden Vergleiche besteht für 104 der 136 möglichen Paarungen der untersuchten Verfahren. Hier sind weitere Simulationen notwendig, um die Verfahren besser miteinander vergleichen zu können und die gefundenen Ergebnisse abzusichern.

Ein Aspekt, der bei der Planung zukünftiger Studien beachtet werden sollte, ist eine ausreichende Anzahl an Simulationswiederholungen. Erst durch genügend Wiederholungen einer Simulation werden die Ergebnisse verlässlich und sind damit überhaupt erst zur Analyse von MD-Verfahren geeignet. Dieser Punkt ist bei 76 der 125 untersuchten Studien nicht beachtet worden, wodurch sie zum Vergleich der MD-Verfahren nicht verwendet werden konnten.

Bei der Betrachtung der untersuchten Verfahren ist auffällig, dass fast alle der in den Studien untersuchten Imputationsverfahren ausschließlich für quantitative Daten geeignet sind oder nur anhand quantitativer Daten untersucht werden. Die Behandlung von fehlenden Werten bei qualitativen Daten wird hingegen weitestgehend vernachlässigt. Hier existiert ein großes Potential für weitere Forschung.

5 Literaturverzeichnis

Acuna, Edgar; Rodriguez, Caroline (2004): The Treatment of Missing Values and its Effect on Classifier Accuracy. In: David Banks, Frederick R. McMorris, Phipps Arabie und Wolfgang Gaul (Hg.): Classification, Clustering, and Data Mining Applications. Berlin, Heidelberg: Springer (Studies in Classification, Data Analysis, and Knowledge Organisation), S. 639–647.

Aittokallio, Tero (2010): Dealing with Missing Values in Large-scale Studies: Microarray Data Imputation and Beyond. In: *Briefings in Bioinformatics* 11 (2), S. 253–264. DOI: 10.1093/bib/bbp059.

Ali, A. M. G.; Dawson, S-J; Blows, F. M.; Provenzano, E.; Ellis, I. O.; Baglietto, L. et al. (2011): Comparison of Methods for Handling Missing Data on Immunohistochemical Markers in Survival Analysis of Breast Cancer. In: *British Journal of Cancer* 104 (4), S. 693–699. DOI: 10.1038/sj.bjc.6606078.

Allison, Paul D. (2001): Missing Data. Thousand Oaks et al.: SAGE (Quantitative Applications in the Social Sciences, 136).

Ambler, Gareth; Omar, Rumana Z.; Royston, Patrick (2007): A Comparison of Imputation Techniques for Handling Missing Predictor Values in a Risk Model with a Binary Outcome. In: *Statistical Methods in Medical Research* 16 (3), S. 277–298. DOI: 10.1177/0962280206074466.

Backhaus, Klaus; Blechschmidt, Boris (2009): Fehlende Werte und Datenqualität. In: *Die Betriebswirtschaft* 69 (2), S. 265–287.

Bankhofer, Udo (1995): Unvollständige Daten- und Distanzmatrizen in der Multivariaten Datenanalyse. Bergisch Gladbach, Köln: Eul (Quantitative Ökonomie, 64).

Barzi, Federica; Woodward, Mark (2004): Imputations of Missing Values in Practice: Results from Imputations of Serum Cholesterol in 28 Cohort Studies. In: *American Journal of Epidemiology* 160 (1), S. 34–45. DOI: 10.1093/aje/kwh175.

Batista, Gustavo E. A. P. A.; Monard, Maria Carolina (2003): An Analysis of Four Missing Data Treatment Methods for Supervised Learning. In: *Applied Artificial Intelligence* 17 (5-6), S. 519–533. DOI: 10.1080/713827181.

Bello, Abdul L. (1993): A Simulation Study of Imputation Techniques in Linear Quadratic and Kernel Discriminant Analyses. In: *Journal of Statistical Computation and Simulation* 48 (3-4), S. 167–180. DOI: 10.1080/00949659308811549.

Bø, Trond Hellem; Dysvik, Bjarte; Jonassen, Inge (2004): LSimpute: Accurate Estimation of Missing Values in Microarray Data with Least Squares Methods. In: *Nucleic Acids Research* 32 (3), S. e34. DOI: 10.1093/nar/gnh026.

Branden, Karlien Vanden; Verboven, Sabine (2009): Robust Data Imputation. In: *Computational Biology and Chemistry* 33 (1), S. 7–13. DOI: 10.1016/j.compbiolchem.2008.07.019.

Brock, Guy N.; Shaffer, John R.; Blakesley, Richard E.; Lotz, Meredith J.; Tseng, George C. (2008): Which Missing Value Imputation Method to Use in Expression Profiles: A Comparative Study and Two Selection Schemes. In: *BMC Bioinformatics* 9 (1). DOI: 10.1186/1471-2105-9-12.

van Buuren, Stef (2012): Flexible Imputation of Missing Data. Boca Raton: Chapman and Hall/CRC (Interdisciplinary Statistics Series).

Chiu, Chia-Chun; Chan, Shih-Yao; Wang, Chung-Ching; Wu, Wei-Sheng (2013): Missing Value Imputation for Microarray Data: A Comprehensive Comparison Study and a Web Tool. In: *BMC Systems Biology* 7(Suppl. 6). DOI: 10.1186/1752-0509-7-S6-S12.

Cox, Brenda G.; Folsom, Ralph E. (1978): An Empirical Investigation of Alternative Item Nonresponse Adjustments. In: American Statistical Association (Hg.): Proceedings of the Section on Survey Research Methods. American Statistical Association. Washington: American Statistical Association, S. 219–223.

Crawford, Sybil L.; Tennstedt, Sharon L.; McKinlay, John B. (1995): A Comparison of Analytic Methods for Non-Random Missingness of Outcome Data. In: *Journal of Clinical Epidemiology* 48 (2), S. 209–219. DOI: 10.1016/0895-4356(94)00124-9.

de Souto, Marcilio C. P.; Jaskowiak, Pablo A.; Costa, Ivan G. (2015): Impact of Missing Data Imputation Methods on Gene Expression Clustering and Classification. In: *BMC Bioinformatics* 16. DOI: 10.1186/s12859-015-0494-3.

Demirtas, Hakan; Hedeker, Donald (2008): Imputing Continuous Data under some Non-Gaussian Distributions. In: *Statistica Neerlandica* 62 (2), S. 193–205. DOI: 10.1111/j.1467-9574.2007.00377.x.

Devi Priya, R.; Sivaraj, R. (2015): A Review of Missing Data Handling Methods. In: *International Journal On Engineering Technology and Sciences* 2 (2), S. 58–68.

Eekhout, Iris; Boer, Michiel R. de; Twisk, Jos W. R.; Vet, Henrica C. W. de; Heymans, Martijn W. (2012): Missing Data. A Systematic Review of How They Are Reported and Handled. In: *Epidemiology* 23 (5), S. 729–732. DOI: 10.1097/EDE.0b013e3182576cdb.

Enders, Craig K. (2010): Applied Missing Data Analysis. New York: Guilford Press (Methodology in the Social Sciences).

Farhangfar, Alireza; Kurgan, Lukasz; Dy, Jennifer (2008): Impact of Imputation of Missing Values on Classification Error for Discrete Data. In: *Pattern Recognition* 41 (12), S. 3692–3705. DOI: 10.1016/j.patcog.2008.05.019.

Feelders, Ad (1999): Handling Missing Data in Trees: Surrogate Splits or Statistical Imputation? In: Jan M. Zytkow und Jan Rauch (Hg.): Principles of Data Mining and Knowledge Discovery. Berlin, Heidelberg: Springer (Lecture Notes in Computer Science, 1704), S. 329–334.

García-Laencina, Pedro J.; Sancho-Gómez, José-Luis; Figueiras-Vidal, Aníbal R. (2010): Pattern Classification with Missing Data: A Review. In: *Neural Comput & Applic* 19 (2), S. 263–282. DOI: 10.1007/s00521-009-0295-6.

Groenwold, Rolf H. H.; Donders, A. Rogier T.; Roes, Kit C. B.; Harrell, Frank E.; Moons, Karel G. M. (2012): Dealing with Missing Outcome Data in Randomized Trials and Observational Studies. In: *American Journal of Epidemiology* 175 (3), S. 210–217. DOI: 10.1093/aje/kwr302.

Hallgren, Kevin A.; Witkiewitz, Katie (2013): Missing Data in Alcohol Clinical Trials: A Comparison of Methods. In: *Alcoholism: Clinical and Experimental Research* 37 (12), S. 2152–2160. DOI: 10.1111/acer.12205.

Ibrahim, Joseph G.; Chen, Ming-Hui; Lipsitz, Stuart R.; Herring, Amy H. (2005): Missing-Data Methods for Generalized Linear Models. A Comparative Review. In: *Journal of the American Statistical Association* 100 (469), S. 332–346. DOI: 10.1198/016214504000001844.

Kim, Hyunsoo; Golub, Gene H.; Park, Haesun (2005): Missing Value Estimation for DNA Microarray Gene Expression Data: Local Least Squares Imputation. In: *Bioinformatics* 21 (2), S. 187–198. DOI: 10.1093/bioinformatics/bth499.

Kim, Hyunsoo; Golub, Gene H.; Park, Haesun (2006): Corrigendum. Missing Value Estimation for DNA Microarray Gene Expression Data: Local Least Squares Imputation. In: *Bioinformatics* 22 (11), S. 1410–1411. DOI: 10.1093/bioinformatics/btk053.

Laaksonen, Seppo (2003): Alternative Imputation Techniques for Complex Metric Variables. In: *Journal of Applied Statistics* 30 (9), S. 1009–1020. DOI: 10.1080/0266476032000076137.

Little, Roderick J. A.; Rubin, Donald B. (2002): Statistical Analysis with Missing Data. 2. Aufl. Hoboken: Wiley (Wiley series in probability and statistics).

Liu, Peng; El-Darzi, Elia; Lei, Lei; Vasilakis, Christos; Chountas, Panagiotis; Huang, Wei (2005): An Analysis of Missing Data Treatment Methods and Their Application to Health Care Dataset. In: Xue Li, Shuliang Wang und Zhao Yang Dong (Hg.): Advanced Data Mining and Applications, Bd. 3584. Berlin, Heidelberg: Springer (Lecture Notes in Computer Science, 3584), S. 583–590.

Luengo, Julián; García, Salvador; Herrera, Francisco (2010): A Study on the Use of Imputation Methods for Experimentation with Radial Basis Function Network Classifiers Handling Missing Attribute Values: The Good Synergy between RBFNs and EventCovering Method. In: *Neural Networks* 23 (3), S. 406–418. DOI: 10.1016/j.neunet.2009.11.014.

Marshall, Andrea; Altman, Douglas G.; Holder, Roger L. (2010a): Comparison of Imputation Methods for Handling Missing Covariate Data when Fitting a Cox Proportional Hazards Model: A Resampling Study. In: *BMC Medical Research Methodology* 10. DOI: 10.1186/1471-2288-10-112.

Marshall, Andrea; Altman, Douglas G.; Royston, Patrick; Holder, Roger L. (2010b): Comparison of Techniques for Handling Missing Covariate Data within Prognostic Modelling Studies: A Simulation Study. In: *BMC Medical Research Methodology* 10. DOI: 10.1186/1471-2288-10-7.

Myrtveit, Ingunn; Stensrud, Erik; Olsson, Ulf H. (2001): Analyzing Data Sets with Missing Data: An Empirical Evaluation of Imputation Methods and Likelihood-Based Methods. In: *IEEE Transactions on Software Engineering* 27 (11), S. 999–1013. DOI: 10.1109/32.965340.

Oh, Sunghee; Kang, Dongwan D.; Brock, Guy N.; Tseng, George C. (2011): Biological Impact of Missing-value Imputation on Downstream Analyses of Gene Expression Profiles. In: *Bioinformatics* 27 (1), S. 78–86. DOI: 10.1093/bioinformatics/btq613.

Penone, Caterina; Davidson, Ana D.; Shoemaker, Kevin T.; Di Marco, Moreno; Rondinini, Carlo; Brooks, Thomas M. et al. (2014): Imputation of Missing Data in Life-History Trait Datasets: Which Approach Performs the Best? In: *Methods Ecol Evol* 5 (9), S. 961–970. DOI: 10.1111/2041-210X.12232.

Pérez, Adriana; Dennis, Rodolfo J.; Gil, Jacky F. A.; Rondón, Martín A.; López, Adriana (2002): Use of the Mean, Hot Deck and Multiple Imputation Techniques to Predict Outcome in Intensive Care Unit Patients in Colombia. In: *Statist. Med.* 21 (24), S. 3885–3896. DOI: 10.1002/sim.1391.

Raymond, Mark R. (1986): Missing Data in Evaluation Research. In: *Evaluation & the Health Professions* 9 (4), S. 395–420. DOI: 10.1177/016327878600900401.

Roth, Philip L. (1994): Missing Data: A Conceptual Review for Applied Psychologists. In: *Personnel Psychology* 47 (3), S. 537–560. DOI: 10.1111/j.1744-6570.1994.tb01736.x.

Roth, Philip L.; Switzer, Fred S.; Switzer, Deborah M. (1999): Missing Data in Multiple Item Scales: A Monte Carlo Analysis of Missing Data Techniques. In: *Organizational Research Methods* 2 (3), S. 211–232. DOI: 10.1177/109442819923001.

Rubin, Donald B. (1996): Multiple Imputation after 18+ Years. In: *Journal of the American Statistical Association* 91 (434), S. 473–489. DOI: 10.1080/01621459.1996.10476908.

Saunders, Jeanne A.; Morrow-Howell, Nancy; Spitznagel, Edward; Doré, Peter; Proctor, Enola K.; Pescarino, Richard (2006): Imputing Missing Data: A Comparison of Methods for Social Work Researchers. In: *Social Work Research* 30 (1), S. 19–31. DOI: 10.1093/swr/30.1.19.

Schafer, Joseph L.; Graham, John W. (2002): Missing Data: Our View of the State of the Art. In: *Psychological Methods* 7 (2), S. 147–177. DOI: 10.1037/1082-989X.7.2.147.

Tsikriktsis, Nikos (2005): A Review of Techniques for Treating Missing Data in OM Survey Research. In: *Journal of Operations Management* 24 (1), S. 53–62. DOI: 10.1016/j.jom.2005.03.001.