Ilmenauer Beiträge zur Wirtschaftsinformatik

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

Steve Röhrig

Eine Simulationsstudie zum Umgang mit fehlenden nominalen Daten

Arbeitsbericht Nr. 2023-01, Dezember 2023

Technische Universität Ilmenau
Fakultät für Wirtschaftswissenschaften
Institut für Wirtschaftsinformatik

Autor: Steve Röhrig

Titel: Eine Simulationsstudie zum Umgang mit fehlenden nominalen Daten

Ilmenauer Beiträge zur Wirtschaftsinformatik Nr. 2023-01, Technische Universität Ilmenau, 2023

ISSN 1861-9223

ISBN 978-3-938940-66-2

urn:nbn:de:gbv:ilm1-2023200311

Anschrift: Technische Universität Ilmenau, Fakultät für Wirtschaftswissenschaften und Medien, Institut für Wirtschaftsinformatik, PF 100565, D-98684 Ilmenau.

Gliederung

Zusammenfassung: Durch eine Imputation lassen sich fehlende Werte innerhalb einer unvollständigen Datenmatrix ersetzen, wodurch eine vollständige Datenmatrix entsteht. Auf dieser Grundlage können anschließend weitere Analysen und Untersuchungen durchgeführt werden. Die Güte der Imputation kann anhand verschiedener Kriterien bewertet werden, die wiederum von den Eigenschaften der Datenmatrix mit fehlenden Werten abhängen. Im Rahmen einer Simulationsstudie werden die Auswirkungen dieser Eigenschaften auf die Ergebnisse verschiedener Imputationsverfahren für nominale Daten untersucht. Neben etablierten Gütekriterien wird auch ein eigens konzipiertes Bewertungskriterium herangezogen, um die Qualität der Imputationen zu beurteilen. Die Ergebnisse dieser Studie verdeutlichen, dass die Wahl eines geeigneten Imputationsverfahrens von den spezifischen Eigenschaften der unvollständigen Datenmatrix sowie den angestrebten Untersuchungszielen abhängt. Ein Verfahren, dass für alle Eigenschaften und Ziele die besten Ergebnisse erzeugt, konnte nicht gefunden werden.

Schlüsselworte: fehlende Werte, Gütekriterien, Imputation, missing data, nominale Daten, qualitative Daten, Simulationsstudie, Verteilungsabweichung

1 Einleitung

Für die Verarbeitung von Daten wird in der Regel eine vollständige Datenmatrix vorausgesetzt. Daher bedarf das Fehlen der Daten einer Behandlung, denn bereits einfachste Rechenoperationen, wie die Bestimmung eines Mittelwertes, sind bei fehlenden Daten nicht möglich. Somit stellen sie für Forschende aus allen Bereichen der Wissenschaft und zahlreiche wirtschaftliche Anwendungen ein Problem dar. Die Auseinandersetzung mit fehlenden Daten bzw. deren Behandlung reicht vom einfachen Löschen der Objekte mit fehlenden Werten bis hin zur Anwendung komplexer Verfahren, um beispielsweise Ersatzwerte, sogenannte Imputationswerte, für die fehlenden Daten zu generieren. Eine Übersicht verschiedener Möglichkeiten zur Behandlung der fehlenden Daten wird beispielsweise von Bankhofer (1995) gegeben. Dabei stellt sich zwangsläufig die Frage, welche Art von Behandlung der fehlenden Daten für eine vorliegende Problemstellung bzw. eine Datenmatrix mit fehlenden Werten zu den besten Ergebnissen führt. In der Literatur existiert keine umfangreiche Handlungsempfehlung zum Umgang mit fehlenden kategorialen Daten, wie sie zum Beispiel Rockel (2022, S. 229-232) für quantitative Merkmale entwickelt hat. Die Bedeutsamkeit dieser Thematik betonen u. a. Cugnata und Salini (2017, S. 316), Ferrari et al. (2011, S. 2410), Rockel (2022, S. 236) sowie Vidotto et al. (2018, S. 56), welche hervorheben, dass die Untersuchung von Verfahren für quantitative Merkmale vorherrschend sei. Deshalb besteht die Notwendigkeit, einen Vergleich von Imputationsverfahren speziell für kategoriale Daten durchzuführen, wie es im Rahmen dieser Simulationsstudie zunächst für nominale Daten gezeigt wird. Darauf aufbauend kann eine entsprechende Handlungsempfehlung zum Umgang mit fehlenden nominalen Daten erarbeitet werden. Im Rahmen dieses Arbeitspapiers werden verschiedene Datenmatrizen und Ausfallszenarien simuliert, anhand derer die Güte verschiedener Imputationsverfahren hinsichtlich unterschiedlicher Kriterien miteinander verglichen wird. Dazu wird im anschließenden Kapitel 2 ein Überblick zum Aufbau und Ablauf der Simulationsstudie gegeben. Die Ergebnisse der so durchgeführten Simulationsstudie werden in Kapitel 3 präsentiert. Im abschließenden Kapitel 4 werden die wichtigsten Erkenntnisse zusammengefasst und ein Ausblick auf weitere Forschungsarbeit gegeben.

2 Aufbau der Simulationsstudie

Im Rahmen dieses Kapitels wird der Aufbau der Simulationsstudie beschrieben. Dabei werden die einzelnen Faktoren, welche im Rahmen der Studie untersucht werden, kurz erläutert. Der Aufbau der Simulationsstudie erfolgt in Anlehnung an die Vorgehensweise zum Vergleich von Missing Data-Verfahren auf Basis von simulierten vollständigen Datenmatrizen (siehe Röhrig und Rockel, 2020, S. 2). Der daraus resultierende Aufbau lässt sich in folgende fünf Abschnitte unterteilen:

1. Datenerzeugung
2. Datenlöschung
3. Datenimputation
4. Datenauswertung
5. Datenverlässlichkeit

Im ersten Schritt erfolgt die Erzeugung bzw. Simulation der vollständigen Daten. Als Parameter werden dazu die Anzahl an Objekten *n*, die Anzahl der Merkmale *m*, die Korrelation zwischen den Merkmalen ρ und die Art der Verteilung festgelegt. Der zweite Schritt umfasst das Löschen der anfänglich erzeugten Daten, abhängig vom gewählten Anteil der fehlenden Daten *p* und dem jeweils festgelegten Ausfallmechanismus. Anschließend erfolgt die Vervollständigung der nun unvollständigen Datenmatrizen mittels Imputationsverfahren. Zur Auswertung findet ein Vergleich der simulierten vollständigen Datenmatrizen A^{sim} und den imputierten vervollständigten Datenmatrizen A^{imp} durch die Betrachtung verschiedener Gütekriterien statt mit

$$A^{sim} = \left(a_{ik}^{sim}\right)_{n\times m} = \begin{pmatrix} a_{11}^{sim} & \cdots & a_{1m}^{sim} \\ \vdots & \ddots & \vdots \\ a_{n1}^{sim} & \cdots & a_{nm}^{sim} \end{pmatrix}.$$

Die Werte welche in der simulierten vollständigen Datenmatrizen A^{sim} gelöscht wurden und damit fehlend sind, werden anschließend imputiert und bilden zusammen mit den vorhandenen Werten von A^{sim} die imputierte vervollständigte Datenmatrix A^{imp}. Im Rahmen der Auswertung wird zusätzlich erfasst, wie sicher die erzielten Ergebnisse sind.

Die Simulationsstudie wurde vollständig mittels des Statistikprogramms R (R Core Team, 2022) in der Version 4.1.3 durchgeführt. Einen Überblick zur Durchführung der Simulation gibt der in Abbildung 1 dargestellte Ablaufplan. Darin werden alle fünf oben genannten Abschnitte zusammen mit den jeweiligen Simulationsparametern abgebildet. Die Abschnitte sind in Abbildung 1 fettgedruckt dargestellt.

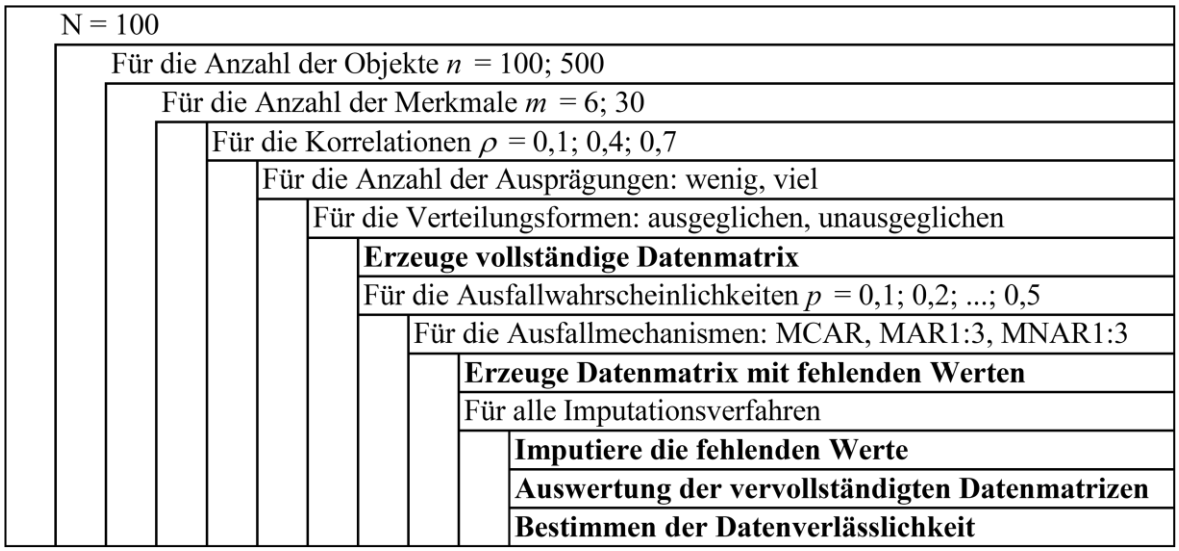

Abbildung 1: Ablaufplan Simulationsstudie

Aus dem Ablaufplan der Simulationsstudie geht hervor, dass die Simulation aller 720 Faktorstufenkombinationen N = 100-mal durchgeführt wird. Dies dient zur Sicherung der Ergebnisse und der damit verbundenen Reliabilität. Der allgemeine Aufbau und Ablauf der Simulationsstudie orientieren sich an Rockel (2022, 171–186), welcher eine Simulationsstudie in ähnlicher Form, jedoch nicht für kategorische Daten, durchgeführt hat. Im Rahmen der vorliegenden Arbeit werden in der Simulationsstudie zum Teil andere Faktorstufen verwendet sowie weitere Anpassungen aufgrund des untersuchten Skalenniveaus durchgeführt. Dazu gehören unter anderem die Verteilungsformen kategorialer Daten sowie die Anwendung verschiedener Gütekriterien, darunter auch ein eigenständig entwickeltes Kriterium. Die einzelnen Stufen des Ablaufplans werden nachfolgend kurz erläutert.

Für die Simulation der vollständigen Daten erfolgt zunächst eine Wahl der Anzahl der Objekte *n* und Anzahl der Merkmale *m* sowie die Korrelationen der Merkmale ρ. Die Auswahl erfolgt analog zu den Ergebnissen von Rockel (2022, S. 172). Demzufolge beträgt die zur Erzeugung der vollständigen Daten verwendete Anzahl der Objekte $n = 100$ und $n = 500$ sowie die Anzahl der Merkmale $m = 6$ und $m = 30$.

Zur Erzeugung der Daten ist es außerdem erforderlich, eine Zielkorrelationsmatrix Σ [1] vorzugeben. Diese hat folgende Form und besitzt abhängig von der Anzahl der Merkmale die Dimensionen $m \times m$:

$$\Sigma = \begin{pmatrix} 1 & \rho & \rho & \dots & \rho \\ \rho & 1 & \rho & \dots & \rho \\ \rho & \rho & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & 1 & \rho \\ \rho & \rho & \cdots & \rho & 1 \end{pmatrix}$$

Als Korrelationsstufen werden $\rho = 0,1; 0,4$ und $0,7$ gewählt, wodurch ein schwacher, mittlerer und starker Zusammenhang zwischen den Merkmalen simuliert wird (vgl. Rockel, 2022, S. 172 f.).

Ein weiterer Untersuchungspunkt ist die Art der Verteilung. Mit dieser wird zum einen die Anzahl der Ausprägungen und zum anderen die Verteilungsform (ausgeglichen oder unausgeglichen) betrachtet, da Untersuchungen ergeben haben, dass eine unausgeglichene beziehungsweise asymmetrische Verteilung einen negativen Einfluss auf die Güte der Imputationsmethoden besitzt (vgl. Rhemtulla et al., 2012, S. 360). Die Art der Verteilung bestimmt im Rahmen dieser Arbeit die Anzahl und die relative Häufigkeit der Ausprägungen a_j^k eines Merkmals *k*, wobei alle möglichen Ausprägungen eines Merkmals *k* in der Ausprägungsmenge A_k zusammengefasst werden. Somit gibt $|A_k|$ die Anzahl der verschiedenen Ausprägungen a_j^k des Merkmals *k* an. In Hinblick auf unterschiedliche Anzahlen von Ausprägungen $|A_k|$ wird im Rahmen dieser Arbeit zwischen Datenmatrizen mit

[1] Die Zielkorrelationsmatrix gibt die gewünschten Korrelationen zwischen den einzelnen Merkmalen vor (vgl. Fialkowski und Tiwari, 2019, S. 2).

wenigen und Datenmatrizen mit vielen Ausprägungen in den einzelnen Merkmalen unterschieden. Konkret besitzt in der Simulationsstudie eine Datenmatrix mit wenigen Ausprägungen zwei, drei oder vier Ausprägungen je Merkmal, daher $|A_k| \in \{2;3;4\}$ für alle $k \in \{1;...;m\}$. Hingegen besitzt eine Datenmatrix mit vielen Ausprägungen entweder fünf, sechs oder sieben Ausprägungen je Merkmal, daher $|A_k| \in \{5;6;7\}$ für alle $k \in \{1;...;m\}$. Hinsichtlich der Art der Verteilung wird im Gegensatz zu Rhemtulla et al. (2012) und daran angelehnten Untersuchungen, wie zum Beispiel Wu et al. (2015), für die Verteilungsformen zwischen ausgeglichen und unausgeglichen verteilten Häufigkeiten der Ausprägungen unterschieden, anstatt asymmetrisch und symmetrisch (vgl. Rhemtulla et al., 2012, S. 360).

Bei einer ausgeglichenen Verteilung liegen ähnlich große Anzahlen an Beobachtungen in den Ausprägungen eines Merkmals vor. Hingegen liegen bei einer unausgeglichenen Verteilung deutlich mehr Beobachtungen in einigen Ausprägungen eines Merkmals vor als in den restlichen (vgl. Agresti, 2019, S. 45). Dazu besitzen im Rahmen dieser Simulationsstudie für den Fall einer ausgeglichenen Verteilung alle Ausprägungen a_j^k eines Merkmals *k* dieselbe relative Häufigkeit $f\left(a_j^k\right) = 1 \div |A_k|$ für $j = 1;...;|A_k|$ und $k = 1;...;m$. Im Fall einer unausgeglichenen Verteilung besitzen in dieser Studie alle bis auf eine Ausprägung a_j^k eines Merkmals *k* eine relative Häufigkeit von $f\left(a_j^k\right) = 0{,}1$ und eine Ausprägung a_j^k eines Merkmals *k* eine relative Häufigkeit 1 minus der Summe der relativen Häufigkeiten aller anderen Ausprägungen des Merkmals *k* und damit:

$$f\left(a_j^k\right) = 1 - \sum_{j=1}^{|A_k|-1} 0{,}1 = 1 - 0{,}1 \cdot \left(|A_k| - 1\right)$$

Als Beispiel wird bei einer unausgeglichenen Verteilung das Merkmal *k* mit $|A_k| = 3$ betrachtet. Die erste und dritte Ausprägung dieses Merkmals besitzen eine relative Häufigkeit von $f\left(a_1^k\right) = f\left(a_3^k\right) = 0{,}1$ und die zweite Ausprägung eine relative Häufigkeit von $f\left(a_2^k\right) = 1 - 0{,}1 \cdot (3 - 1) = 0{,}8$.

Anhand dieser Parameter werden die simulierten vollständigen Datenmatrizen A^{sim} mit der Funktion corrvar2() aus dem R-Paket SimCorrMix (Fialkowski, 2022) in der Version 0.1.1 erzeugt. Eine genaue Beschreibung der Funktionsweise ist Fialkowski und Tiwari (2019) zu entnehmen. Die so generierten Daten sind zunächst ordinalskaliert und werden anschließend in nominale Daten transformiert.[2] In vergleichbaren Studien variierten die angewandten Ausfallraten in den betrachteten Simulationsstudien in einem Bereich von etwa 10 % bis 50 %. Beispielsweise wurde dies im Rahmen der Literaturrecherche von Röhrig und Rockel (2020, S. 6 ff.) bzw. Rockel (2022, S. 135) oder Lin und Tsai (2020, S. 1490–1491) festgestellt. Deshalb werden für diese Simulationsstudie die Abstufungen der Ausfallwahrscheinlichkeiten von $p = 0{,}1; 0{,}2; 0{,}3; 0{,}4$ und 0,5 gewählt.

Als Ausfallmechanismen werden neben dem Missing at Randon (MAR)- und dem Missing Completely at Random (MCAR)-Ausfallmechanismus, wie sie auch überwiegend in den bereits existierenden Simulationsstudien untersucht wurden (vgl. Röhrig und Rockel, 2020, S. 7; Rockel, 2022, S. 135), auch der Missing Not at Random (MNAR)-Ausfallmechanismus betrachtet. Eine allgemeine Definition der verschiedenen Ausfallmechanismen ist beispielsweise Little und Rubin (2020, S. 13–23) zu entnehmen. Für den in dieser Arbeit verwendeten MCAR-Ausfallmechanismus wird stets der Anteil p der Beobachtungswerte in den vom Ausfall betroffenen Merkmalen zufällig gelöscht. Vom Ausfall betroffen sind im Rahmen dieser Simulationsstudie alle Merkmale mit gerader Indexnummer $k = 2; 4; ...; m$. Somit werden in jedem zweiten Merkmal Werte gelöscht.

Der verwendete MAR-Ausfallmechanismus löscht ebenso wie der MCAR-Ausfallmechanismus auch in jedem zweiten Merkmal bzw. in allen Merkmalen mit gerader Indexnummer zufällig Werte (vgl. Rockel, 2022, S. 173–175). Der Ausfall wird dabei von dem jeweiligen Vorgänger des vom Ausfall betroffenen Merkmals bestimmt. Demzufolge werden die Merkmale $k = 1; 3; ...; m - 1$ als ausfallsteuernde Merkmale bezeichnet.

[2] Aufgrund der Skalendegression geht die annäherungsweise bestimmte Zielkorrelation zwischen den ordinalen Merkmalen verloren, diese kann jedoch erneut ermittelt werden. Da im späteren Verlauf der Arbeit lediglich die Abweichungen der Korrelationen untersucht werden, ist dies unkritisch.

Die Objekte des ausfallsteuernden Merkmals werden in zwei Gruppen unterteilt, wobei die Ausfallwahrscheinlichkeit in den vom Ausfall betroffenen Merkmalen dreimal so hoch ist, je nach Gruppenzugehörigkeit des Objekts im jeweiligen ausfallsteuernden Merkmal. Wenn also ein Objekt in einem ausfallsteuernden Merkmal der Gruppe mit der höheren Ausfallwahrscheinlichkeit angehört, wird der Wert dieses Objekts im dazugehörigen vom ausfallbetroffenen Merkmal mit einer dreimal so hohen Wahrscheinlichkeit gelöscht, als es die Objekte aus der anderen Gruppe werden. Daraus resultiert der MAR1:3-Ausfallmechanismus, wobei in der Bezeichnung das Verhältnis dieser Ausfallwahrscheinlichkeiten angegeben ist.

Der MNAR-Ausfallmechanismus verhält sich gleich zum MAR-Ausfallmechanismus, mit dem einzigen Unterschied, dass der Ausfall der Daten nicht mehr vom Vorgängermerkmal gesteuert wird, sondern vom Ausfall betroffenen Merkmal selbst. Daher werden die Objekte in dem vom Ausfall betroffenen Merkmal wiederum in zwei Gruppen unterteilt, wobei die Ausfallwahrscheinlichkeit auch hier je nach Gruppenzugehörigkeit dreimal so hoch ist wie in der anderen Gruppe. Damit wird der MNAR1:3-Ausfallmechanismus verwendet. Sowohl beim MCAR- und MAR- als auch beim MNAR-Ausfallmechanismus werden die Daten stets so gelöscht, dass in den vom Ausfall betroffenen Merkmalen insgesamt $n \cdot p$ Werte gelöscht werden. Eine exakte Definition der verwendeten Ausfallmechanismen ist Rockel (2022, S. 173–175) zu entnehmen. Das Löschen der Daten erfolgt mittels des R-Pakets missMethods von Rockel (2020) in der Version 0.3.0.

Als Imputationsverfahren wird eine Auswahl der von Röhrig und Rockel (2020, S. 10) identifizierten Verfahren betrachtet. Eine Übersicht der betrachteten Verfahren ist gemeinsam mit dem zur Anwendung des Verfahrens verwendeten R-Pakets sowie der Versionsnummer in der folgenden Tabelle 1 angegeben. Außerdem befinden sich in der Tabelle Hinweise zur weiterführenden Literatur, in der die Funktionsweise der einzelnen Verfahren erklärt wird, da eine entsprechende Darstellung weit über den Rahmen dieses Arbeitspapieres hinausgehen würde. Mittels der in Tabelle 1 gezeigten Imputationsverfahren wird für jede Faktorstufenkombination die imputierte Datenmatrix A^{imp} 100-mal erzeugt.

Zum Vergleich dieser Imputationsverfahren werden vier verschiedene Gütekriterien betrachtet, welche nachfolgend kurz erläutert werden. Die Auswahl der Gütekriterien orientiert sich dabei an den von Röhrig und Rockel (2020, S. 10–11) gefundenen Vergleichskriterien.

Imputationsverfahren	**R-Paket (Version) & Quelle**	**Weitere Literatur**
EM kategorisch	cat (0.0-7) (Harding et al., 2012)	Schafer (1997, S. 260–264)
EM (stetig) deterministisch mit anschließendem Runden	missMethods (0.3.0) (Rockel, 2020)	Rockel (2022, S. 35–39)
EM (stetig) stochastisch mit anschließendem Runden	missMethods (0.3.0) (Rockel, 2020)	Rockel (2022, S. 35–39)
Entscheidungsbäume (missForest)	missForest (1.5) (Stekhoven, 2022)	Stekhoven und Bühlmann (2012)
Iterative robust model-based imputation (irmi)	VIM (6.1.1) (Templ et al., 2021)	Kowarik und Templ (2016, S. 9–11)
Modusimputation (Modus)	Eigene Umsetzung	Bankhofer (1995, S. 106–108)
Multiple Korrespondenzanalyse (MCA)	missMDA (1.18) (Husson und Josse, 2020)	Josse und Husson (2016)
Nächste-Nachbarn-Verfahren (kNN)	VIM (6.1.1) (Templ et al., 2021)	Troyanskaya et al. (2001, S. 521)
Random Hot-Deck (HD)	VIM (6.1.1) (Templ et al., 2021)	Andridge und Little (2010)

Tabelle 1: Übersicht betrachteter Imputationsverfahren

Das erste Gütekriterium ist die proportion of falsely classified entries (PFC), welche von Stekhoven und Bühlmann (2012, S. 113–114) beschrieben und in einer modifizierten Form angewendet wird. Damit ist es möglich, die Ungenauigkeit der Imputationswerte zu beurteilen, indem der Anteil falsch imputierter Kategorien bzw. Ausprägungen ermittelt wird.

Zur Berechnung des modifizierten PFC[3] wird die Anzahl der nicht übereinstimmenden Imputationswerte a_{ik}^{imp} mit den Werten der simulierten vollständigen Datenmatrix a_{ik}^{sim} bestimmt und mittels Anzahl der Objekte *n* normiert:

$$PFC = \frac{\sum_{i=1}^{n}\sum_{k=1}^{m} \mathrm{I}}{n \cdot m}$$

Dabei nimmt die Indikatorvariable I den Wert 1 an, falls $a_{ik}^{imp} \neq a_{ik}^{sim}$ und sonst den Wert 0 für $i = 1;...;n$ und $k = 1;...;m$. Im Gegensatz zu Stekhoven und Bühlmann (2012) wird hier die Anzahl der Falschklassifizierungen nicht durch die Anzahl der fehlenden Werte geteilt, sondern durch die Anzahl aller Werte.[4]

Als zweites Gütekriterium erfolgt ein Vergleich der Verteilung der Daten vor und nach der Imputation. Dieses Gütekriterium ist eine neuartige und eigenständige Konzeption und wurde demzufolge in den zuvor untersuchten Simulationsstudien in dieser Form nicht betrachtet (vgl. Röhrig und Rockel, 2020, S. 10–11). Hierbei werden jedoch erhebliche Abweichungen in den Ergebnissen zwischen den Imputationsverfahren vermutet. Für die Betrachtung werden die absoluten Häufigkeiten $h\left(a_j^k\right)$ bzw. relativen Häufigkeiten $f\left(a_j^k\right)$ der Ausprägungen a_j^k in den Merkmalen *k* herangezogen. Zur Bestimmung der Abweichung der Verteilung in einem Merkmal muss zunächst eine Häufigkeitstabelle für jedes Merkmal *k* bestimmt werden. Dies wird sowohl für die simulierte vollständige Datenmatrix A^{sim} als auch für die imputierte vervollständigte Datenmatrix A^{imp} durchgeführt. Dadurch ergeben sich für die vollständige Datenmatrix A^{sim} die absoluten Häufigkeiten $h\left(a_j^{k,sim}\right)$ und für die imputierte vervollständigte Datenmatrix A^{imp} die absoluten Häufigkeiten $h\left(a_j^{k,imp}\right)$ für $j = 1;...;\left|A_k\right|$ und $k = 1;...;m$. Anschließend sind die betragsmäßigen Abweichungen der absoluten Häufigkeiten der Ausprägungen $h\left(a_j^{k,sim}\right)$ eines Merkmals *k* zwischen den simulierten Daten und den imputierten Daten $h\left(a_j^{k,imp}\right)$ zu ermitteln.

[3] Im Folgenden aus Gründen der Übersichtlichkeit PFC genannt.

[4] Offensichtlich wird dieses Gütekriterium vom Anteil der fehlenden Werte beeinflusst, jedoch ist diese Abhängigkeit gewollt und Teil der Untersuchung.

Die theoretisch maximal mögliche Abweichung, welche hinsichtlich der absoluten Häufigkeiten in einem Merkmal auftreten kann, beträgt genau zweimal der Anzahl der Objekte *n*. Dies resultiert aus der Überlegung, dass die Summe der absoluten Häufigkeiten in einem Merkmal stets der Anzahl der Objekte *n* entspricht. Das gilt sowohl für die Häufigkeiten $h\left(a_j^{k,sim}\right)$ der simulierten Daten als auch für die der imputierten Daten $h\left(a_j^{k,imp}\right)$ dieser Simulationsstudie. Somit addiert sich die maximale Abweichung zwischen den Häufigkeiten zu 2*n*, was anhand der nachfolgenden Berechnung für ein einziges Merkmal *k* demonstriert wird:

$$\sum_{j=1}^{|A_k|}\left|h\left(a_j^{k,imp}\right)-h\left(a_j^{k,sim}\right)\right| \le \sum_{j=1}^{|A_k|}h\left(a_j^{k,imp}\right)+h\left(a_j^{k,sim}\right) = \sum_{j=1}^{|A_k|}h\left(a_j^{k,imp}\right)+\sum_{j=1}^{|A_k|}h\left(a_j^{k,sim}\right) = n+n = 2n$$

Als Beispiel sind nachfolgend zwei Häufigkeitstabellen in Abbildung 2 dargestellt, wobei die absoluten Häufigkeiten der einzelnen Ausprägungen eines Merkmals *k* einer Datenmatrix vor dem Löschen und nach der Imputation angegeben sind.

$a_j^{k,sim}$	1	2	3
$h\left(a_j^{k,sim}\right)$	9	10	11

$\Rightarrow$ Löschen, Imputation

$a_j^{k,imp}$	1	2	3
$h\left(a_j^{k,imp}\right)$	12	8	10

Abbildung 2: Beispiel Abweichung absolute Häufigkeiten

In diesem Beispiel beträgt die theoretisch maximal mögliche Abweichung $30 \cdot 2 = 60$. Die tatsächliche Abweichung ermittelt sich im Beispiel aus $|12-9|+|8-10|+|10-11| = 6,$ womit sich eine Abweichung der Verteilung in dem betrachteten Merkmal von $6 \div 60 = 0{,}1$ ergibt. Anzumerken ist dabei, dass der Wertebereich des mittels 2*n* normierten Quotienten nicht von 0 bis 1 reicht, sondern maximal den Wert

$$1-\frac{\min_j\left(h\left(a_j^{k,sim}\right)\right)}{n}$$

annimmt.[5] Dies resultiert aus der Tatsache, dass die Häufigkeiten aller Ausprägungen $h\left(a_j^{k,sim}\right)$, welche in der simulierten vollständigen Datenmatrix vorliegen, stets größer 0

[5] Der maximale Wert reduziert sich auf den Anteil der gelöschten Werte *p*, sofern der dargestellte maximale Werte nicht kleiner als *p* ist. Dies ist durch einfaches Nachrechnen nachvollziehbar.

sind, was durch die zuvor festgelegte Anzahl der Ausprägungen und der Verteilungsform bedingt ist. Anstatt mittels 2*n* zu normieren, führt eine Normierung mittels des folgenden Ausdrucks zu einem Wertebereich von 0 bis 1:

$$\left\lfloor n - \min_j \left(h\left(a_j^{k,sim} \right) \right) \right\rfloor \cdot 2$$

Das Zutreffen dieser Normierung wird nachfolgend anhand des Beispiels aus der Abbildung 2 verdeutlicht. Die Verteilung der imputierten Daten, welche für das Beispiel aus Abbildung 2 am denkbar ungünstigsten wäre, ist die folgende:

$a_j^{k,imp}$	1	2	3
$h\left(a_j^{k,imp}\right)$	30	0	0

Für diese Verteilung ergibt sich eine Abweichung von $|30-9|+|0-10|+|0-11|=42$. Diese maximale Abweichung von 42 ist mit der zuletzt dargestellten Normierung direkt zu berechnen: $(30-9)\cdot 2=42$. Somit würde eine Normierung mit diesem Wert zu einem Wertebereich von 0 bis 1 führen. Jedoch hätte die Anwendung dieser Normierung zur Folge, dass für unterschiedliche minimale Häufigkeiten bei gleichbleibenden Differenzen zwischen den Häufigkeitstabellen unterschiedliche Abweichungen entstünden. Für die dargestellte denkbar ungünstige Verteilung für das Beispiel aus Abbildung 2 ergibt sich mittels einer Normierung durch 2*n* eine Abweichung von $42 \div 60 = 0,7.$ Das entspricht genau dem bereits beschriebenen maximalen Wert von:

$$1-\frac{9}{30}=0,7\,.$$

Ein Vorteil beider Normierungen ist, dass diese auch bei Multiplikation der Häufigkeiten in beiden Tabellen mit beliebigen positiven Zahlen dieselbe Abweichung wie zuvor ergeben.

Um eine bessere Vergleichbarkeit der Ergebnisse und damit genauere Rückschlüsse auf das Beibehalten der ursprünglichen Verteilung ziehen zu können, wird im Rahmen der Simulation auf die Normierung mittels $2n$ zurückgegriffen.[6]

Diese Betrachtung wird für alle Merkmale durchgeführt, aufsummiert und anschließend durch die Anzahl der vom Ausfall betroffenen Merkmale m_{mis} geteilt, woraus sich die mittlere Abweichung der Verteilung in den Merkmalen mit fehlenden Werten ergibt. Diese Berechnung kann alternativ mit den relativen Häufigkeiten der einzelnen Ausprägungen $f\left(a_j^{k,sim}\right)$ bzw. $f\left(a_j^{k,imp}\right)$ durchgeführt werden. Die genauen Formeln zur Berechnung der mittleren Verteilungsabweichung VA lauten damit:

$$VA = \frac{1}{m_{mis}} \sum_{k=1}^{m} \frac{\sum_{j=1}^{|A_k|} \left| h\left(a_j^{k,imp}\right) - h\left(a_j^{k,sim}\right) \right|}{2n} = \frac{1}{m_{mis}} \sum_{k=1}^{m} \frac{\sum_{j=1}^{|A_k|} \left| f\left(a_j^{k,imp}\right) - f\left(a_j^{k,sim}\right) \right|}{2}$$

Als drittes Gütekriterium wird überprüft, inwiefern der Zusammenhang zwischen 2 Merkmalen nach der Imputation wiederhergestellt werden kann. Dies wird anhand von Cramér's V verglichen. Dazu wird für jedes Merkmalpaar Cramér's V der simulierten Datenmatrix A^{sim} und der imputierten Datenmatrix A^{imp} bestimmt. Für die Berechnung von Cramér's V wird die Funktion `assocstats()` des R-Pakets vcd (Meyer et al., 2022) verwendet. Die auf diese Weise bestimmten paarweise Korrelationen können in der Korrelationsmatrix C^{sim} und C^{imp} zusammengefasst werden. Anschließend kann auf Grundlage dieser die Wurzel aus den mittleren quadratischen Abweichungen RMSE zwischen den Korrelationen der simulierten vollständigen Daten c_{kl}^{sim} und denen der imputierten Daten c_{kl}^{imp} bestimmt werden. Die zugrundeliegende Berechnungsvorschrift ergibt sich demnach folgendermaßen:[7]

$$RMSE_C = \sqrt{\frac{2}{m \cdot (m-1)} \sum_{k=1}^{m} \sum_{l>k} \left(c_{kl}^{imp} - c_{kl}^{sim} \right)^2}$$

[6] In der Realität ist es denkbar, dass beispielsweise im Rahmen einer Umfrage eine Antwortmöglichkeit von keiner der befragten Personen gewählt wird und damit eine Häufigkeit von 0 besitzt. In dieser Situation schöpft eine Normierung mittels $2n$ den Wertebereich zwischen 0 und 1 vollständig aus, wobei die Alternative Normierung bei einer minimalen Häufigkeit von 0 auch zu einer Normierung mittels $2n$ führt.

[7] Dabei ist zu beachten, dass die Korrelation eines Merkmals zu sich selbst stets 1 und damit kein Gegenstand dieser Untersuchung ist.

Als viertes und letztes Gütekriterium werden die Auswirkungen der Imputationsverfahren auf eine logistische Regression untersucht, welches das am meisten genutzte Verfahren im Fall einer kategorialen abhängigen Variable ist (vgl. Agresti, 2013, S. 163). Im Gegensatz zu den zuvor beschriebenen Gütekriterien, werden an dieser Stelle nur die Ergebnisse für $n = 500$ betrachtet. Dies ergibt sich aus der Tatsache, dass eine hohe Anzahl von Regressionskoeffizienten geschätzt werden muss. Zur Durchführung der logistischen Regression wurde eine zusätzliche Variable *y* der Datenmatrix hinzugefügt, welche die abhängige Variable der logistischen Regression darstellt. Diese Variable ist binomialverteilt mit $p = 0{,}5$ sowie $n = 500$ und damit $Y \sim B(500;0{,}5)$. Die Auswirkung auf eine logistische Regression wird anhand der RMSE zwischen den Regressionskoeffizienten $\beta_j^{k,sim}$ der simulierten vollständigen Datenmatrix A^{sim} und den Regressionskoeffizienten $\hat{\beta}_j^{k,imp}$ der imputierten Datenmatrix A^{imp} für alle $k \in \{1;...;\, m\}$ und $j \in \{2;...;\, |A_k|\}$ bestimmt.[8] Das Modell der logistischen Regression besitzt abhängig von der jeweiligen Art der Verteilung die folgende Struktur, wobei auf der linken Seite der folgenden Gleichung die logarithmierten Odds, auch Logit genannt, der abhängigen Variable *y* dargestellt sind:

$$\log\left(\frac{P(Y=1\mid X)}{1-P(Y=1\mid X)}\right) = \beta_0 + \sum_{k=1}^{m}\sum_{j=2}^{|A_k|}\beta_j^k \cdot x_j^k$$

Zu beachten ist, dass die logistische Regression die abhängige Variable *y* nicht konkret modelliert, sondern die Wahrscheinlichkeit untersucht, dass diese einer bestimmten Ausprägung ($Y = 1$) gegeben den unabhängigen Variablen *X* angehört (vgl. James et al., 2017, S. 130). Dabei handelt es sich bei den im Modell dargestellten x_j^k um Indikatorvariablen, welche anzeigen, ob ein Objekt in einem Merkmal *k* die Ausprägung a_j^k besitzt ($x_j^k = 1$) oder nicht ($x_j^k = 0$). Als Beispiel ergibt sich für eine Datenmatrix mit 4 Merkmalen sowie $|A_1| = |A_4| = 2$ und $|A_2| = |A_3| = 3$, das folgende Modell:

$$\log\left(\frac{P(Y=1\mid X)}{1-P(Y=1\mid X)}\right) = \beta_0 + \beta_2^1 \cdot x_2^1 + \beta_2^2 \cdot x_2^2 + \beta_3^2 \cdot x_3^2 + \beta_2^3 \cdot x_2^3 + \beta_3^3 \cdot x_3^3 + \beta_2^4 \cdot x_2^4$$

[8] $j \in \{2,..., \|A_k\|\}$ resultiert aus der Tatsache, dass in R als Referenzkategorie in den einzelnen Merkmalen die erste Ausprägung gewählt wird (ähnlich einer Dummy-Codierung), weshalb sich die Anzahl der Regressionskoeffizienten pro Merkmal hinsichtlich der Ausprägungen um eins reduziert (vgl. Agresti, 2019, S. 100, R Core Team, 2022).

Die Anzahl der Regressionskoeffizienten lässt sich demnach folgendermaßen berechnen:

$$\sum_{k=1}^{m} |A_k| - m + 1$$

Im Beispiel ergibt dies: $2+3+3+2-4+1=7$ Regressionskoeffizienten. Zur Berechnung der RMSE zwischen den Regressionskoeffizienten wird die folgende Gleichung verwendet:

$$RMSE_{\beta} = \sqrt{\frac{1}{\sum_{k=1}^{m} |A_k| - m + 1} \left(\left(\hat{\beta}_0^{imp} - \beta_0^{sim} \right)^2 + \sum_{k=1}^{m} \sum_{j=2}^{|A_k|} \left(\hat{\beta}_j^{k,imp} - \beta_j^{k,sim} \right)^2 \right)}$$

Die Regressionskoeffizienten $\beta_j^{k,sim}$ und $\hat{\beta}_j^{k,imp}$ der simulierten vollständigen Datenmatrix A^{sim} bzw. der imputierten Datenmatrix A^{imp} sowie die jeweiligen Absolutglieder werden mittels der im Statistikprogramm R standardmäßig implementierten Funktion `glm()` mit dem Attribut `family = binomial()` berechnet.

Im Rahmen der Auswertung wird auch die Reliabilität der Daten bzw. Ergebnisse betrachtet. Ein Maß, um die Reliabilität der Ergebnisse genauer zu beurteilen, ist der Monte-Carlo-Standardfehler $\hat{\sigma}_{\mathrm{MC}}$, welchen Morris et al. (2019) zum Quantifizieren der Unsicherheit der Simulation vorschlagen. Der Monte-Carlo-Standardfehler $\hat{\sigma}_{\mathrm{MC}}$ berechnet sich unter den gegebenen Voraussetzungen folgendermaßen (vgl. Morris et al., 2019, S. 2086):

$$\hat{\sigma}_{\mathrm{MC}} = \sqrt{\frac{1}{N(N-1)} \sum_{i=1}^{N} \left(\theta_i - \bar{\theta} \right)^2}$$

Dabei ist *N* die Anzahl der Wiederholungen, θ_i der in der Wiederholung *i* berechnete Wert für ein Gütekriterium eines Verfahrens bei festgelegter Faktorstufenkombination und $\bar{\theta}$ der Mittelwert aus den θ_i über alle Wiederholungen *N*. Damit ist es möglich, die Unsicherheit der Simulationsstudie bei einer festgelegten Faktorstufenkombination für ein Verfahren hinsichtlich eines Gütekriteriums zu beurteilen. Je kleiner der Monte-Carlo-Standardfehler $\hat{\sigma}_{\mathrm{MC}}$ ist, desto sicherer sind die gewonnenen Ergebnisse. Anhand der Formel ist zudem erkennbar, dass eine Erhöhung der Anzahl der Wiederholungen *N* einen po-

sitiven Effekt auf die Reliabilität der Ergebnisse besitzt. Die Berechnung des Monte-Carlo-Standardfehlers $\hat{\sigma}_{\text{MC}}$ erfolgt somit für jede Kombination aus Verfahren, Faktorstufenkombination sowie Gütekriterium und damit $9 \cdot 720 \cdot 4 = 25920$ mal. Als Beispiel besitzt ein betrachtetes Imputationsverfahren bei einer bestimmten Faktorstufenkombination und 8 durchgeführten Wiederholungen die folgenden Werte bezüglich eines Gütekriteriums:

i	1	2	3	4	5	6	7	8
θ_i	0,06	0,049	0,052	0,052	0,053	0,051	0,051	0,051

Damit ergibt sich der Mittelwert als $\bar{\theta} = (0,06 + 0,049 + ... + 0,051) \div 8 \approx 0,052$ und der daraus resultierende Monte-Carlo-Standardfehler mit

$$\hat{\sigma}_{\text{MC}} = \sqrt{\frac{1}{8(8-1)} \cdot \left[(0,06 - 0,052)^2 + (0,049 - 0,052)^2 + ... + (0,051 - 0,052)^2 \right]} \approx 0,001 .$$

3 Ergebnisse der Simulationsstudie

Die Beurteilung der durch die Simulationsstudie erzeugten Ergebnisse erfolgt unterteilt nach den zuvor beschriebenen Gütekriterien. Hierfür werden die Mittelwerte eines Gütekriteriums über alle Wiederholungen für jede Faktorstufenkombination und jedes Verfahren berechnet. Es ergeben sich insgesamt 25920 Datenpunkte. Um eine übersichtlichere Darstellung zu gewährleisten, sind die erzielten Werte nach den Ergebnissen mit wenigen Objekten (n = 100) und denen mit vielen Objekten (n = 500) unterteilt.[9] Für n = 100 und für n = 500 sind die entsprechenden Ergebnisse in den nachfolgenden Abschnitten in den Abbildungen 3 bis 9 dargestellt. Diese weisen einen nahezu identischen Aufbau auf, sodass an dieser Stelle eine allgemeine Erläuterung erfolgt. In den ersten drei Spalten sind die Ergebnisse hinsichtlich des MCAR-Ausfallmechanismus dargestellt. Diese drei Spalten sind wiederum unterteilt in die Ergebnisse für $\rho = 0,1; 0,4$ und $0,7$. Gleiches gilt für die Spalten vier bis sechs und sieben bis neun, wobei hier die Ergebnisse für die Ausfallmechanismen MAR bzw. MNAR zu sehen sind. In den Zeilen befinden sich die verschiedenen Ver-

[9] Dies trifft nicht für die Auswertung der logistischen Regression zu. Hier werden die Ergebnisse nur für n = 500 betrachtet.

teilungsformen sowie die Anzahl an Merkmalen. Die Verteilungsform gibt an, wie viele Ausprägungen verwendet wurden und welche Art der Verteilung zugrunde liegt. Die Verteilungsform „wenig“ bezieht sich auf Datenmatrizen mit zwei, drei oder vier Ausprägungen je Merkmal, daher $|A_k| \in \{2;3;4\}$ für alle $k \in \{1;...;m\}$. Die Verteilungsform „viel“ hingegen bezieht sich auf Datenmatrizen mit fünf, sechs oder sieben Ausprägungen je Merkmal, also $|A_k| \in \{5;6;7\}$ für alle $k \in \{1;...;m\}$. Der Ausdruck „gleich“ signalisiert eine ausgeglichene bzw. symmetrische Verteilung bei der ähnlich große Häufigkeiten der Beobachtungen in den Ausprägungen eines Merkmals vorliegen. Die Bezeichnung „ungleich“ beschreibt den Umstand, dass in einer Ausprägung eines Merkmals deutlich mehr Beobachtungen vorliegen als in den restlichen Ausprägungen (vgl. Agresti, 2019, S. 45). Die betrachtete Anzahl an Merkmalen mit $m = 6$ bzw. $m = 30$ befindet sich ebenfalls in den Zeilen. Die Ergebnisse bezüglich einer Verteilungsform sind dabei zunächst für $m = 6$ und anschließend für $m = 30$ dargestellt. Außerdem ist auf der Abszisse für alle Kombinationen der zuvor benannten Faktoren der Anteil der fehlenden Daten abzulesen und auf der Ordinate die jeweiligen Ergebnisse für das betrachtete Gütekriterium.

Für $n = 100$ wird zusätzlich in den entsprechenden Abschnitten eine Tabelle angegeben, um einen besseren Vergleich der Auswirkungen der unterschiedlichen Faktorstufen zu ermöglichen. In diesen Tabellen wird für alle Teilabbildungen[10] die maximale Spannweite (max SP) zwischen dem besten und dem schlechtesten Ergebnis einer Ausfallrate sowie der Mittelwert der besten Ergebnisse (best mean) über die fünf Ausfallraten gebildet werden. Der Aufbau dieser Tabellen ist identisch zu denen der Abbildungen 3 bis 9, wobei anstelle der Ausfallraten und Ergebnisse nun die eben beschriebenen maximalen Spannweiten und Mittelwerte dargestellt sind. Zusätzlich werden für eine bessere Einordnung der Ergebnisse der einzelnen Verfahren in den entsprechenden Abschnitten außerdem für alle Datenpunkte Ränge hinsichtlich aller Verfahren vergeben. Die empirische Verteilungs-

[10] Eine Teilabbildung stellt eine Faktorstufenkombination mit Ausnahme der Ausfallrate dar. Eine Teilabbildung besteht somit aus 5 Datenpunkten eines jeden Verfahrens (sofern vorhanden).

funktion $F(x)$ dieser Ränge wird für alle Verfahren einer dazugehörigen Tabelle angegeben (auf 3 Nachkommastellen gerundet).

Für $n = 500$ wird betrachtet, inwiefern sich die Ergebnisse aufgrund der Erhöhung der Anzahl der Objekte verändern. Dies basiert unter anderem auf einem Vergleich der jeweiligen Abbildungen für $n = 100$ und für $n = 500$. Außerdem werden die maximalen Verbesserungen (max) und Verschlechterungen (min) sowie mittlere Abweichungen (mean) in einer Tabelle angegeben. Im Gegensatz zu $n = 100$ werden für $n = 500$ nicht die empirischen Verteilungen der Ränge der einzelnen Verfahren betrachtet, sondern die Änderungen, welche sich hinsichtlich der relativen Häufigkeit $f(x)$ der einzelnen Ränge aufgrund der Erhöhung der Anzahl der Objekte für jedes Verfahren ergeben. Dazu werden die relativen Häufigkeiten der Ränge der einzelnen Verfahren für $n = 100$ von denen für $n = 500$ subtrahiert. Das Ergebnis sind Abweichungen, wobei positive Werte bedeuten, dass ein Verfahren häufiger einen Rang unter $n = 500$ erreicht hat als unter $n = 100$. Negative Werte beschreiben den umgekehrten Fall. Diese Tabellen können neben den Einträgen in den Zellen selbst auch spalten- bzw. zeilenweise interpretiert werden. Die Spalten stellen die Veränderungen der relativen Häufigkeiten der Ränge des jeweiligen Verfahrens dar und innerhalb der Zeilen ist erkennbar, wie sich die Vergabe der Ränge zwischen den Verfahren verschoben hat.[11] Abschließend werden die Ergebnisse für jedes Gütekriterium kurz zusammengefasst. Eine Besonderheit, welche in den Abbildungen 3 bis 9 der folgenden Abschnitte auffällt, besteht darin, dass die Ergebnisse für die EM-Imputation speziell für kategoriale Daten nur für $m = 6$ zu sehen sind. Dies liegt an dem Umstand, dass die genutzte Implementierung des Verfahrens die Imputationen nur für eine begrenzte Gesamtzahl an Kategorien durchführen kann.[12]

[11] Die Abweichungen in Zeilen und Spalten ergeben in Summe 0, wobei geringfügige Abweichungen aufgrund von Rundungsfehlern möglich sind.

[12] Dies ging aus der Korrespondenz mit einem der Entwickler des R-Pakets hervor. Eine Alternative Implementierung ist nicht verfügbar.

3.1 Evaluation des Anteils falsch imputierter Werte

In diesem Kapitel werden die Ergebnisse bezüglich des Anteils der falsch imputierten Werte (PFC) evaluiert. Die Konstruktion des Gütekriteriums legt nahe, dass ein Verfahren umso besser abschneidet, je niedriger sein PFC ist. Im Folgenden werden die Ergebnisse hinsichtlich des PFC für $n = 100$ betrachtet und damit die Abbildung 3. Die Abbildung zeigt deutlich, dass mit zunehmender Ausfallrate die Ergebnisse aller Verfahren wie erwartet schlechter werden. Außerdem weichen die Ergebnisse unterschiedlich stark voneinander ab, je nach zugrundeliegender Art der Verteilung. Besonders bei den gleichmäßigen Verteilungen mit vielen Ausprägungen sind die Ergebnisse der einzelnen Verfahren sehr ähnlich. Des Weiteren fällt auf, dass mit steigender Korrelation auch der Unterschied zwischen den einzelnen Verfahren zunimmt. Dies ist für nahezu alle Verteilungsformen und Ausfallmechanismen zu beobachten. Zusätzlich ist zu sehen, dass die Ergebnisse der besten Verfahren schlechter beim MNAR-Ausfallmechanismus sind. Nachfolgend sollen Besonderheiten einzelner Verfahren hervorgehoben werden, welche in Abbildung 3 erkennbar sind. Es wird deutlich, dass MCA überwiegend zu den besten Verfahren gehört, mit Ausnahme des MNAR-Ausfallmechanismus und vereinzelt für den MAR-Ausfallmechanismus. MissForest erzielt besonders gute Ergebnisse, wenn die Anzahl der Merkmale $m = 30$ ist. Schlechter werden sie hingegen, wenn die Verteilung unausgeglichen ist und die Anzahl der Merkmale $m = 6$ beträgt. Irmi schneidet in den meisten Fällen nur etwas schlechter als die Gruppe der besten Verfahren ab. Dementsprechend ist ein besonders guter Ergebnisbereich auch nicht hervorzuheben. Die Ergebnisse der Imputation mittels kNN befinden sich überwiegend im Mittelfeld. Für Random Hot Deck existiert keine besonders geeignete Faktorstufenkombination. Die EM-Imputation speziell für kategoriale Daten ist das Verfahren, welches für ungleiche Verteilungen die mit Abstand schlechtesten Ergebnisse erzeugt. Unter dem MNAR-Ausfallmechanismus liefert sie jedoch verhältnismäßig gute Werte. Im Vergleich sind die Ergebnisse der deterministischen EM-Imputation gut, wenn wenige Ausprägungen vorliegen und die Anzahl der Merkmale $m = 6$ ist. Schlechter werden die Ergebnisse hingegen bei ungleichen Verteilungen und $m = 30$.

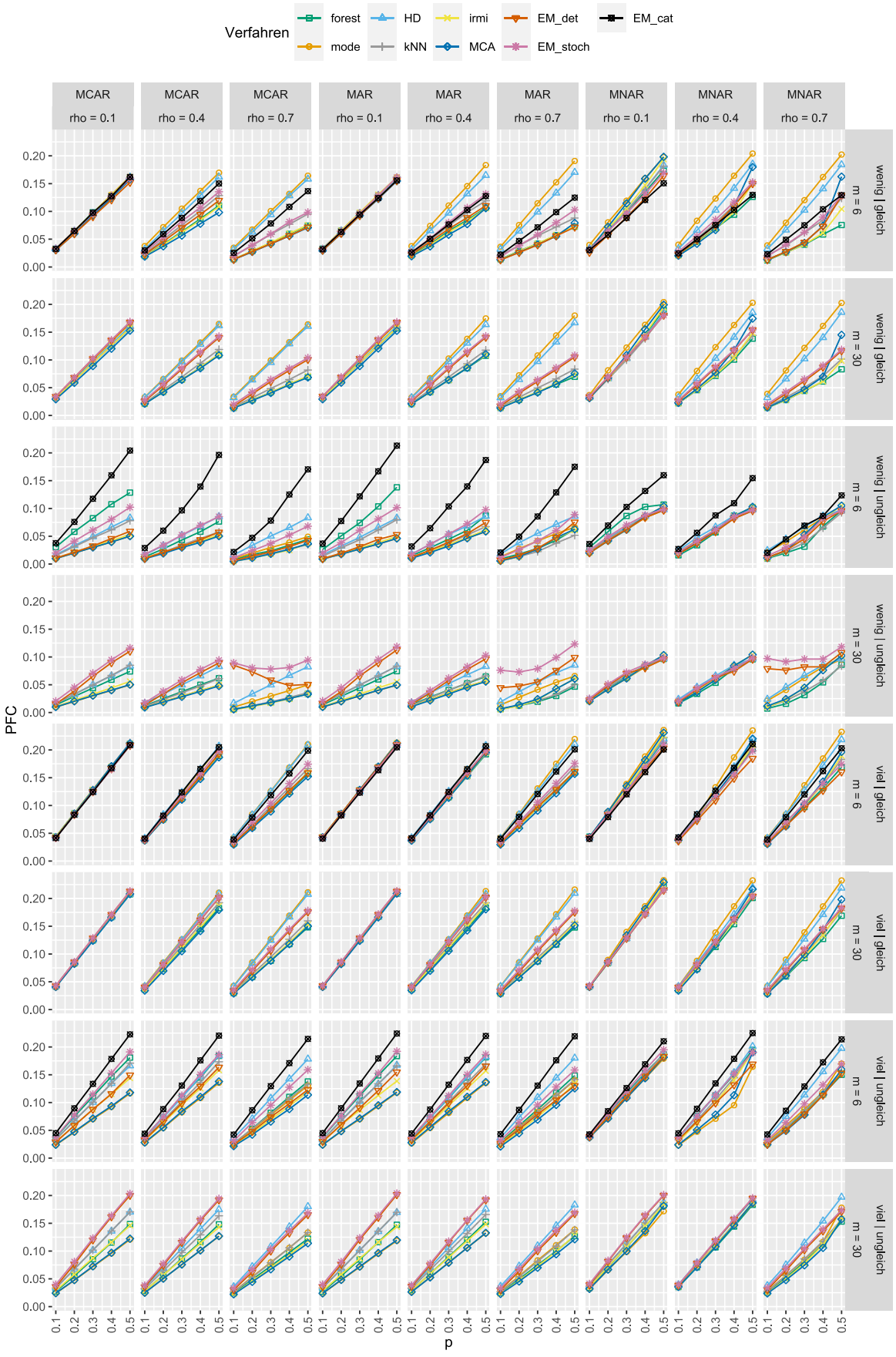

Abbildung 3: Anteil der falsch imputierten Werte bei $n = 100$

Für die stochastische EM-Imputation existiert ebenfalls keine sonderlich geeignete Faktorstufenkombination. Bei einer ungleichen Verteilung und $m = 30$ sind die Ergebnisse dieser vergleichsweise am schlechtesten. Die Modusimputation besitzt eine besondere Stellung, da sie aufgrund der ungleichen Verteilung stark begünstigt wird, falls der ursprüngliche Modus gelöscht würde. Das erklärt die guten Werte für ungleiche Verteilungen besonders dann, wenn in der Datenmatrix viele Ausprägungen vorliegen. Für den Fall einer ausgeglichenen Verteilung liefert die Modusimputation entsprechend schlechte Ergebnisse. Trotz der günstigen Situation einer ungleichen Verteilung kann sich die Modusimputation nicht von den besten Verfahren absetzen.

In der nachfolgenden Tabelle werden die Ergebnisse aus Abbildung 3 in Form von Verteilungsparametern aggregiert. Dazu sind die maximalen Spannweiten und die Mittelwerte der besten Ergebnisse einer jeden Teilabbildung angegeben.

	MCAR rho = 0,1	MCAR rho = 0,4	MCAR rho = 0,7	MAR rho = 0,1	MAR rho = 0,4	MAR rho = 0,7	MNAR rho = 0,1	MNAR rho = 0,4	MNAR rho = 0,7		
max SP	0,0106	0,0714	0,0935	0,0079	0,0781	0,1191	0,0475	0,0779	0,1268	m = 6	w | g
best mean	0,0907	0,0577	0,0417	0,0919	0,0594	0,0411	0,0886	0,0698	0,0424		
max SP	0,016	0,0566	0,0956	0,0161	0,0672	0,1101	0,0248	0,0646	0,1194	m = 30	w | g
best mean	0,09	0,0641	0,0409	0,0901	0,0639	0,0416	0,103	0,0756	0,0461		
max SP	0,154	0,1462	0,1354	0,1671	0,1288	0,1235	0,0631	0,0589	0,0378	m = 6	w | n
best mean	0,0299	0,0294	0,0187	0,0271	0,0336	0,026	0,0607	0,0568	0,0438		
max SP	0,0664	0,0466	0,0847	0,0694	0,0472	0,0767	0,0109	0,0124	0,0899	m = 30	w | n
best mean	0,0299	0,0284	0,0181	0,0298	0,0334	0,0227	0,0599	0,0547	0,0386		
max SP	0,0055	0,02	0,0579	0,0085	0,0159	0,0627	0,035	0,0504	0,0724	m = 6	v | g
best mean	0,1244	0,1113	0,0904	0,1228	0,1142	0,0917	0,12	0,1103	0,0951		
max SP	0,0059	0,031	0,0631	0,0058	0,0326	0,0682	0,0183	0,0321	0,0635	m = 30	v | g
best mean	0,124	0,106	0,0882	0,1241	0,1066	0,0876	0,1277	0,1148	0,0954		
max SP	0,1061	0,0849	0,1007	0,1065	0,0846	0,0937	0,0312	0,0835	0,0638	m = 6	v | n
best mean	0,0698	0,0815	0,0663	0,0707	0,0812	0,0712	0,1078	0,0802	0,0829		
max SP	0,0825	0,0678	0,0661	0,0853	0,0613	0,0617	0,0325	0,0139	0,0485	m = 30	v | n
best mean	0,0719	0,0757	0,0677	0,0712	0,0791	0,0707	0,1001	0,108	0,081		

Tabelle 2: Maximale Spannweite und Mittelwerte für PFC und $n = 100$

In Tabelle 2 wird deutlich, dass für alle Verteilungsformen mit einer ausgeglichenen Verteilung die maximalen Spannweiten zwischen den Ergebnissen der Verfahren mit steigender Korrelation zunehmen und damit einen größeren Unterschied zueinander aufweisen, was bereits rein optisch in Abbildung 3 zu erkennen ist. Für fast alle Faktorstufenkombina-

tionen sinkt mit steigender Korrelation außerdem der Mittelwert der besten Ergebnisse. Eine Ausnahme davon bilden die ungleichen Verteilungen in Kombination mit dem MAR-Ausfallmechanismus, hier weichen die Mittelwerte davon ab. Dies ist auch für die unausgeglichene Verteilung mit vielen Ausprägungen und $m = 6$ unter dem MNAR-Ausfallmechanismus sowie für alle unausgeglichenen Verteilungen mit vielen Ausprägungen und $m = 30$ der Fall. Anhand der Tabelle wird außerdem ersichtlich, dass die Verfahren unter dem MNAR-Ausfallmechanismus größere Spannweiten aufweisen als unter dem MCAR- und MAR-Ausfallmechanismus, wobei es für die Verteilung mit vielen Ausprägungen zu Ausnahmen kommt, da hier die besten Verfahren unter MNAR schlechter abschneiden, was die Spannweite reduziert. Eine weitere Ausnahme davon ist für die ungleiche Verteilung mit wenigen Ausprägungen zu finden. Hier ist die Spannweite unter MNAR deutlich geringer, was lediglich daran liegt, dass die EM-Imputation für kategoriale Daten verhältnismäßig gute Ergebnisse liefert. Bezüglich der besten Ergebnisse sind diese ähnlich oder etwas schlechter unter dem MNAR-Ausfallmechanismus als die des MCAR- und MAR-Ausfallmechanismus. Hinsichtlich der Erhöhung der Merkmalsanzahl sind bezüglich der Spannweite geringe Reduzierungen zu erkennen, wohingegen auf die Mittelwerte der besten Ergebnisse keine Effekte zu sehen sind. Die unterschiedlichen Verteilungen führen zu deutlichen Unterschieden in den maximalen Spannweiten der Ergebnisse. Auch diese Beobachtung ist bereits rein optisch in Abbildung 3 erkennbar. So sind die maximalen Spannweiten deutlich größer für wenige Ausprägungen als für viele. Außerdem führen ausgeglichene Verteilungen zu höheren Spannweiten als ungleiche Verteilungen. Analog dazu verhalten sich die Mittelwerte der besten Verfahren. Diese sind für den Fall weniger Ausprägungen besser als im Fall vieler Ausprägungen. Auch hier erzeugen die unausgeglichenen Verteilungen bessere Ergebnisse als die ausgeglichenen.

In der nachfolgenden Tabelle 3 sind zusätzlich die empirischen Verteilungsfunktionen der Ränge bezüglich des PFCs der betrachteten Verfahren dargestellt.

$F(x)$	forest	mode	HD	kNN	irmi	MCA	EM det	EM stoch	EM cat
≤1	0.172	0.214	0.008	0.033	0.039	0.444	0.072	0.008	0.039
≤2	0.358	0.275	0.033	0.108	0.253	0.717	0.172	0.033	0.053
≤3	0.544	0.317	0.067	0.2	0.664	0.797	0.264	0.081	0.072
≤4	0.689	0.356	0.1	0.514	0.822	0.856	0.486	0.106	0.075
≤5	0.767	0.461	0.203	0.875	0.892	0.883	0.694	0.147	0.086
≤6	0.864	0.55	0.431	0.961	0.931	0.925	0.781	0.447	0.111
≤7	0.925	0.6	0.683	0.972	0.978	0.969	0.997	0.664	0.222
≤8	0.983	0.825	0.969	0.994	0.989	0.994	1	0.997	0.247
≤9	1	1	1	1	1	1	1	1	1

Tabelle 3: Empirische Verteilungsfunktion der Ränge für PFC und $n = 100$

Zu den besten Verfahren gehört überwiegend MCA. Tatsächlich ist es in ca. 44 % der Fälle das beste Verfahren und zählt in ca. 80 % der Fälle zu den besten drei Verfahren. Die Imputation mittels missForest zählt mit MCA am häufigsten zu den besten zwei Verfahren. In ca. 36 % der Fälle trifft dies zu und für ca. 54 % zählt es zu den drei besten Verfahren. Ähnliche Ergebnisse kann irmi erzielen, in ca. 25 % der Fälle gehört es zu den zwei besten Verfahren und in ca. 66 % zu den besten drei. Vergleichsweise schlechte Ergebnisse können mit irmi nicht beobachtet werden, da es zu ca. 82 % unter den besten vier Verfahren liegt. Die Ergebnisse der Imputation mittels kNN befinden sich überwiegend im Mittelfeld. Dies ist in Tabelle 3 daran erkennbar, dass dieses Verfahren zu ca. 87,5 % unter den besten fünf liegt, wobei 36,1 % davon auf den Rang 5 und 31,4 % auf Rang 4 entfallen. Random Hot-Deck ist zu 79,7 % Rang 6 oder schlechter, wobei am häufigsten Rang 8 mit 28,6 % auftritt. Demzufolge existiert keine Faktorstufenkombination, für die Random Hot-Deck besonders geeignet wäre. Die EM-Imputation speziell für kategorische Daten ist in 50 % der Fälle auf Rang 9, da für $m = 30$ keine Ergebnisse vorhanden sind. Demzufolge verbleiben weitere 25,3 % auf Rang 9 auf Basis der Ergebnisse für $m = 6$, wobei weitere 11,1 % auf Rang 7 entfallen. Aufgrund des schlechten Abschneidens sind auch hier keine Faktorstufenkombinationen besonders positiv hervorzuheben. Die deterministische EM-Imputation hat durchwachsene Ergebnisse erzielt, wobei ein Großteil in den Rängen 1 bis 5 liegt (69,4 %). Die Ergebnisse der stochastischen EM-Imputation sind vergleichsweise we-

niger gut. In 85,3 % aller Fälle befindet diese sich zwischen Rang 6 und 9. Demzufolge existiert auch hier keine Faktorstufenkombination, für welche die stochastische EM-Imputation besonders geeignet wäre. Die Modusimputation ist trotz der bereits beschriebenen besonderen Stellung nur in 31,7 % der Ergebnisse unter den besten 3 Verfahren, obwohl in der Hälfte der Untersuchung eine ungleiche Verteilung vorliegt.

Die Ergebnisse hinsichtlich des PFC für $n = 500$ sind in Abbildung 4 dargestellt. Wie aus dem Vergleich der Abbildung 3 und der Abbildung 4 hervorgeht, verändern sich die Ergebnisse nur geringfügig. In wenigen Ausnahmefällen können Verfahren von der Erhöhung der Anzahl der Objekte profitieren bzw. sich durch diese verschlechtern. Besonders für den Fall einer ungleichen Verteilung wirkt sich die Erhöhung der Anzahl der Objekte für die deterministische EM-Imputation positiv aus. Außerdem können verhältnismäßig gute Ergebnisse für den Fall des MNAR-Ausfallmechanismus beobachtet werden.

Die geringfügigen Veränderungen der Ergebnisse durch die Erhöhung der Anzahl der Objekte werden insbesondere deutlich, wenn die nachfolgende Tabelle 4 betrachtet wird. In dieser sind die maximalen Verbesserungen (positive Werte) und Verschlechterungen (negative Werte) jedes Verfahrens sowie die durchschnittlichen Veränderungen, welche durch die Erhöhung der Anzahl der Objekte erreicht werden, dargestellt.

	forest	mode	HD	kNN	irmi	MCA	EM_det	EM_stoch	EM_cat
min	-0,024	-0,054	-0,012	-0,019	-0,028	-0,036	-0,038	-0,011	-0,03
max	0,02	0,059	0,009	0,02	0,032	0,05	0,054	0,045	0,021
mean	-0,002	0,003	-0,001	0	0,002	0,007	0,008	0,017	-0,005

Tabelle 4: Veränderung der Ergebnisse des PFC durch die Erhöhung von n

Aus Tabelle 4 geht anhand der geringen positiven (max) und negativen (min) Abweichungen sowie den mittleren (mean) Abweichungen nahe 0 hervor, dass sich die Ergebnisse für den PFC mit $n = 500$ ähnlich verhalten, wie die, für $n = 100$. Am stärksten können im Schnitt sowohl die deterministische als auch die stochastische EM-Imputation sowie MCA von der Erhöhung der Anzahl der Objekte profitieren.

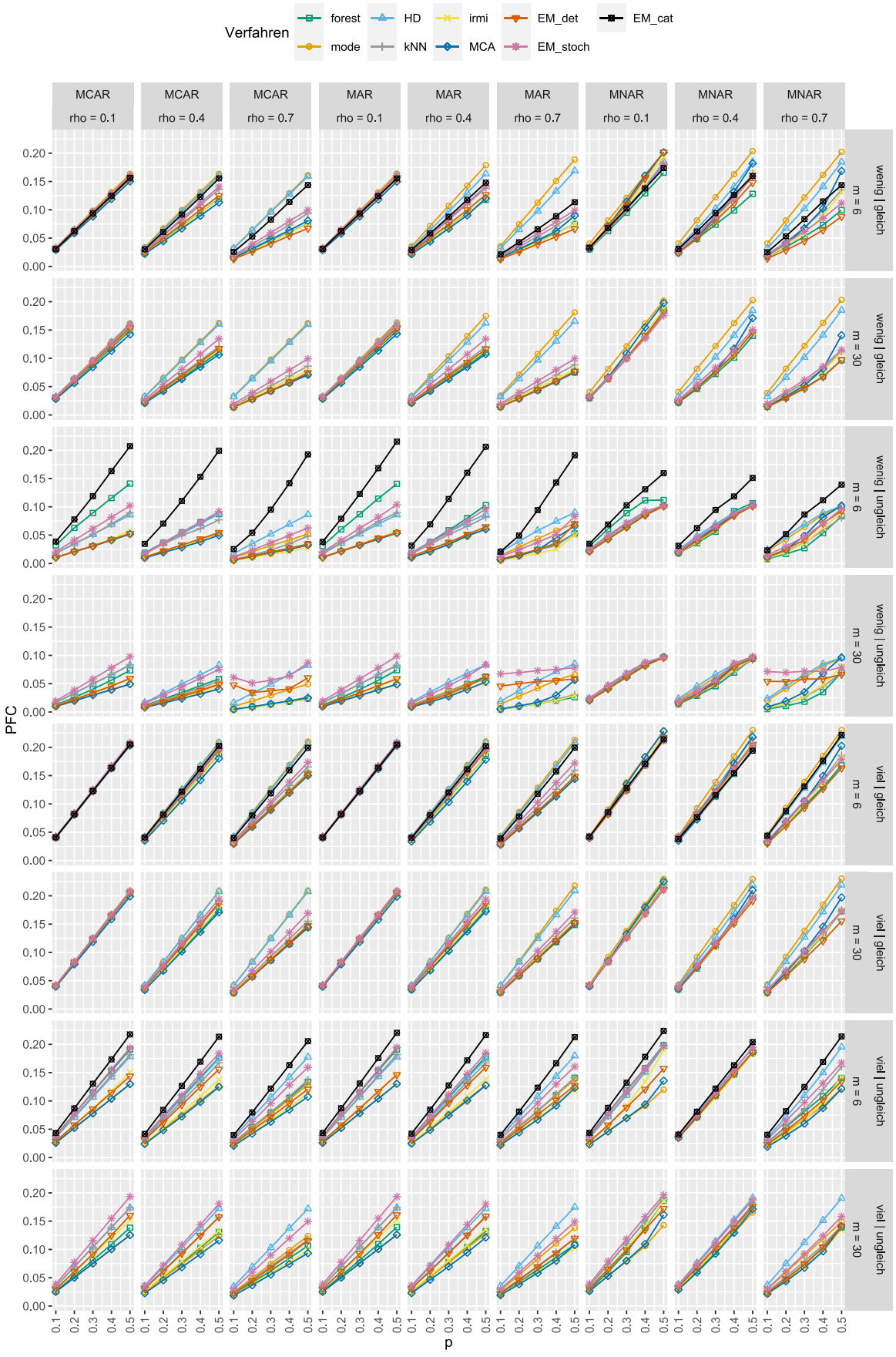

Abbildung 4: Anteil der falsch imputierten Werte bei $n = 500$

Dies zeigt sich auch in der nachfolgenden Tabelle 5, in welcher die Abweichungen der relativen Häufigkeiten der Ränge aller Verfahren für $n = 500$ zu $n = 100$ dargestellt sind.

f(x)	forest	mode	HD	kNN	irmi	MCA	EM det	EM stoch	EM cat
1	-0,025	-0,056	-0,008	-0,025	0,033	0,075	0,050	0,006	-0,036
2	0,084	0,017	-0,025	-0,014	0,070	-0,081	-0,050	-0,011	0,005
3	0,047	-0,009	-0,028	-0,030	-0,120	-0,030	0,183	-0,026	0,003
4	-0,092	0,011	-0,008	-0,181	0,112	0,036	0,103	0,003	0,017
5	0,036	-0,060	-0,015	0,086	-0,020	0,011	-0,052	0	0,011
6	-0,056	0,014	-0,066	0,145	-0,014	0	-0,057	0,053	-0,014
7	-0,002	0,011	0,092	0,030	-0,042	-0,005	-0,180	0,058	0,025
8	0,025	0,025	0,086	-0,019	-0,008	0	0,003	-0,094	-0,005
9	-0,017	0,047	-0,028	0,008	-0,011	-0,006	0	0,011	-0,006

Tabelle 5: Abweichung der relativen Häufigkeiten der Ränge für PFC

In Tabelle 5 wird ersichtlich, dass nur minimale Verschiebungen in den Ranghäufigkeiten auftreten. Die größten Verbesserungen in Bezug auf Rangplätze zeigen sich bei der deterministischen EM-Imputation und missForest. Mit einer Erhöhung der Objektanzahl platziert sich missForest insgesamt häufiger unter den besten drei Verfahren, während die deterministische EM-Imputation wesentlich häufiger zu den besten vier Verfahren gehört. MCA zeigt mit einer erhöhten Objektanzahl häufiger die beste Leistung und irmi erscheint nun häufiger unter den beiden besten Verfahren. Modusimputation, Random Hot Deck und kNN verlieren ihre Platzierungen in den obersten vier Rängen und verschieben sich stattdessen in niedrigere Rangbereiche. Die Platzierungen der stochastischen EM-Imputation und der speziell für kategoriale Daten ändern sich kaum.

Zusammenfassend hat die Anzahl der Objekte nur einen geringen Einfluss auf die Ergebnisse der betrachteten Verfahren. Insbesondere die deterministische EM-Imputation zeigt eine erkennbare Verbesserung bei einer höheren Anzahl von Objekten. Die Erhöhung der Anzahl der Merkmale wirkt sich unterschiedlich auf die Verfahren aus. Es kann kein Effekt für die EM-Imputation speziell für kategoriale Daten beobachtet werden, da hier nur Ergebnisse für $m = 6$ vorliegen. Lediglich Random Hot-Deck und irmi erfahren eine geringe Veränderung der Ergebnisse. Positive und negative Auswirkungen besitzt die Erhöhung der Anzahl der Merkmale für MCA. Im Fall weniger Objekte führt die höhere Anzahl der

Merkmale zu einer Verschlechterung der Ergebnisse für die deterministische und stochastische EM-Imputation. Für bestimmte Faktorstufenkombinationen können missForest, die Modusimputation und kNN von der Erhöhung der Merkmalsanzahl profitieren. Eine deutliche Verschlechterung liefert erwartungsgemäß der Anteil fehlender Werte p, wobei alle Verfahren ähnlich stark von der Verschlechterung betroffen sind. Eine Erhöhung des Zusammenhangs (ρ) besitzt hingegen einen positiven Effekt auf die Ergebnisse fast aller Verfahren. Einzig die Modusimputation und Random Hot-Deck bleiben, wie zu erwarten, von diesem Effekt unberührt. Die Ergebnisse unter dem MCAR- und MAR-Ausfallmechanismus sind sich sehr ähnlich, weshalb hier kein besonderer Effekt beobachtet werden kann. Der MNAR-Ausfallmechanismus führt im Vergleich zu den beiden anderen Ausfallmechanismen zu keiner Verbesserung. Generell führt der MNAR-Ausfallmechanismus entweder zu einer geringfügigen Veränderung oder einer deutlichen Verschlechterung der Ergebnisse. Beispiele für letztere sind die Ergebnisse der Imputation mittels MCA, irmi oder Random Hot-Deck. Im Gegensatz dazu verschlechtern sich die Ergebnisse von missForest in geringerem Maße, weshalb missForest häufig am besten agiert. Auch die Ergebnisse für alle drei Formen der EM-Imputation bleiben überwiegend konstant. Hinsichtlich der Verteilungsform erzielen fast alle Verfahren die mit Abstand besten Ergebnisse, wenn es wenige Ausprägungen und eine unausgeglichene Verteilung gibt. Die zweitbesten Ergebnisse werden zum Großteil bei einer gleichmäßigen Verteilung mit wiederum wenigen Ausprägungen erzielt, worauf die ungleiche Verteilung mit vielen Ausprägungen folgt. Die schlechtesten Ergebnisse werden bei einer gleichmäßigen Verteilung mit vielen Ausprägungen erzielt. Eine Ausnahme von dieser Reihenfolge tritt nur bei der Modusimputation auf (hier zeigen unausgeglichene Verteilungen die besten Ergebnisse) sowie bei der EM-Imputation speziell für kategoriale Daten, bei welcher die besten Ergebnisse für den Fall weniger Ausprägungen in Kombination mit einer gleichmäßigen Verteilung erzielt werden. Die deterministische und die stochastische EM-Imputation erzielen bei einer geringen Anzahl an Ausprägungen und einer ausgeglichenen Verteilung bei $m = 30$ und $\rho = 0,7$ vergleichsweise schlechte Ergebnisse.

Es ist jedoch erwähnenswert, dass sich die Güte hinsichtlich des PFC für die deterministische EM-Imputation unter MCAR bei $m = 30$ und $\rho = 0,7$ sogar bei steigender Ausfallrate verbessert, worauf später genauer eingegangen wird.

3.2 Evaluation der Verteilungsabweichung

Die Auswirkungen auf die Verteilungsabweichung (VA) sind in Abbildung 5 und Abbildung 6 dargestellt. Auch hier verdeutlicht ein niedrigerer Wert die bessere Leistung eines Verfahrens. Im Folgenden werden die Unterschiede in den Ergebnissen bezüglich der Verteilungsabweichung für $n = 100$ und damit die Abbildung 5 näher betrachtet.

Es ist zu erkennen, dass auch für die Verteilungsabweichung die Ergebnisse mit zunehmender Ausfallrate schlechter werden. Ebenso haben die Ausfallmechanismen einen deutlichen Einfluss auf die Qualität der Ergebnisse. Sowohl der MCAR- als auch der MAR-Ausfallmechanismus zeigen ähnliche Ergebnisse, während diese unter dem MNAR-Ausfallmechanismus deutlich schlechter ausfallen können. Deutlichere Veränderungen zeigen sich auch hinsichtlich der verwendeten Verteilungsformen. So sind die Ergebnisse bei ungleichen Verteilungen besser als im Vergleich zu ausgeglichenen, ebenso wie bei Verteilungen mit wenigen Ausprägungen im Vergleich zu solchen mit vielen. Besonderheiten einzelner Verfahren werden nachfolgend genauer betrachtet. MissForest gehört zu den besseren Verfahren, zeigt jedoch Schwächen, insbesondere bei $m = 6$ und generell in Verbindung mit der ungleichen Verteilung mit wenigen Ausprägungen bei den Ausfallmechanismen MCAR und MAR. Ähnlich wie missForest erzielt auch kNN durchweg gute Ergebnisse und gehört fast immer zu den besten Verfahren. Keine validen Ergebnisse erzielen die deterministische und die stochachstische EM-Imputation für den MNAR-Ausfallmechanismus bei wenigen Ausprägungen und einer ungleichen Verteilung sowie $p \geq 0,4$. Dies ist daran erkennbar, dass die Verteilungsabweichung hier größer als die verwendete Ausfallrate ist. Trotz dieser Ausnahmen gehören beide fast immer zu den besten Verfahren. Ebenso ist Random Hot Deck häufig Teil der besten Verfahren. Allerdings erzielt es vergleichsweise schlechte Werte bei einer gleichmäßigen Verteilung mit wenigen Ausprägungen für den MAR- und MNAR-Ausfallmechanismus, wenn $\rho = 0,7$ ist.

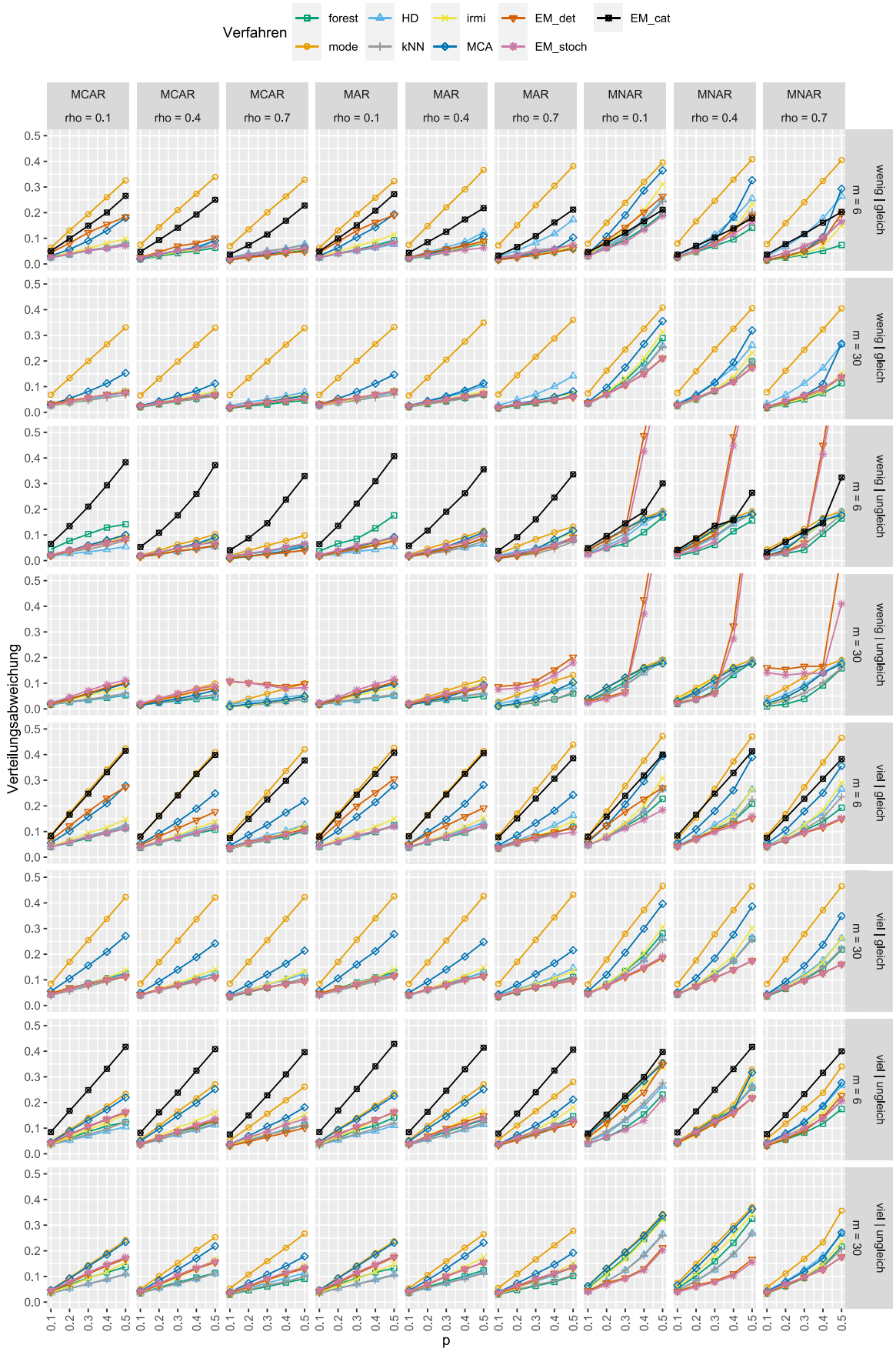

Abbildung 5: Verteilungsabweichung bei $n = 100$

Wie aus der Abbildung ersichtlich ist, befinden sich die Ergebnisse von irmi in den meisten Fällen entweder nah bei den Ergebnissen der besten Verfahren oder zählen selbst zu diesen. MCA zählt häufig zu den schlechteren Verfahren. Im Vergleich zur Gruppe der besten Verfahren fällt seine Leistung deutlich schlechter aus, insbesondere wenn viele Ausprägungen oder eine ausgeglichene Verteilung vorliegen. Sowohl die EM-Imputation speziell für kategoriale Daten als auch die Modusimputation können die ursprüngliche Verteilung am schlechtesten wiederherstellen.

In der nachfolgenden Tabelle 6 werden die maximalen Spannweiten und die Mittelwerte der besten Ergebnisse für die Verteilungsabweichung und $n = 100$ dargestellt. Die Tabelle zeigt, dass die Korrelation nur einen geringen Einfluss auf die maximalen Spannweiten besitzt.

	MCAR rho = 0,1	MCAR rho = 0,4	MCAR rho = 0,7	MAR rho = 0,1	MAR rho = 0,4	MAR rho = 0,7	MNAR rho = 0,1	MNAR rho = 0,4	MNAR rho = 0,7		
max SP	0,1272	0,138	0,1398	0,1231	0,1518	0,1616	0,1038	0,1335	0,1656	m = 6	w | g
best mean	0,0245	0,0205	0,0166	0,0255	0,0215	0,018	0,0496	0,0379	0,02		
max SP	0,1316	0,1332	0,1412	0,1318	0,1414	0,1505	0,0997	0,1165	0,1461	m = 30	w | g
best mean	0,0232	0,0209	0,0156	0,0234	0,0214	0,0181	0,0564	0,0448	0,0286		
max SP	0,1647	0,1591	0,1448	0,1763	0,1465	0,1313	0,3428	0,3198	0,3554	m = 6	w | u
best mean	0,0175	0,0172	0,0119	0,0177	0,0189	0,0174	0,0416	0,0385	0,0352		
max SP	0,0299	0,0267	0,0509	0,0325	0,0321	0,0709	0,3393	0,3479	0,2251	m = 30	w | u
best mean	0,0174	0,0153	0,0109	0,0172	0,0168	0,0146	0,0438	0,0422	0,0316		
max SP	0,1566	0,1509	0,1594	0,1537	0,1467	0,1707	0,1438	0,1589	0,1582	m = 6	v | g
best mean	0,0375	0,0368	0,0334	0,0396	0,0389	0,0342	0,0563	0,0479	0,0465		
max SP	0,1553	0,1557	0,164	0,1559	0,1568	0,1674	0,1402	0,1453	0,1522	m = 30	v | g
best mean	0,0376	0,0376	0,0332	0,0381	0,0383	0,0336	0,056	0,0537	0,0483		
max SP	0,1558	0,1479	0,1478	0,1592	0,1499	0,1443	0,0917	0,0995	0,1125	m = 6	v | u
best mean	0,0351	0,0372	0,0327	0,036	0,0371	0,0374	0,054	0,0605	0,0461		
max SP	0,0667	0,0706	0,088	0,066	0,0758	0,0875	0,0705	0,1072	0,0901	m = 30	v | u
best mean	0,0354	0,0362	0,0303	0,0348	0,0368	0,0321	0,0526	0,0434	0,0493		

Tabelle 6: Maximale Spannweite und Mittelwerte für VA und $n = 100$

Überwiegend führt die Erhöhung der Korrelation zu einem geringfügigen Anstieg der maximalen Spannweite. Mit dieser geht außerdem eine überwiegend geringe Verbesserung der Mittelwerte der besten Verfahren einher. Die Ausfallmechanismen beeinflussen die maximalen Spannweiten kaum, lediglich für die ungleiche Verteilung mit wenigen Ausprägungen sind diese unter MNAR deutlich höher. Dies wird durch die schlechten Ergeb-

nisse der deterministischen und stochastischen EM-Imputation in Abbildung 5 deutlich. Auch die Mittelwerte der besten Verfahren sind unter dem MNAR-Ausfallmechanismus deutlich schlechter als unter MCAR oder MAR. Die Anzahl der Merkmale hat kaum Einfluss auf die maximale Spannweite und die Mittelwerte der besten Ergebnisse, abgesehen von den ungleichen Verteilungen. Hier führt eine Erhöhung der Merkmalsanzahl zu einer Verringerung der maximalen Spannweite und einer Verbesserung der Mittelwerte der besten Verfahren. Wie bereits rein optisch zu erkennen ist, sind die Mittelwerte der besten Verfahren mit wenigen Ausnahmen bei ungleichen Verteilungen besser als bei ausgeglichenen Verteilungen sowie mit wenigen Ausprägungen besser als mit vielen Ausprägungen. Besonders hohe maximale Spannweiten zeigen sich bei ungleichen Verteilungen mit wenigen Ausprägungen unter dem MNAR-Ausfallmechanismus. Für $m = 30$ sind die maximalen Spannweiten für die ungleiche Verteilung fast immer niedriger als für die ausgeglichene Verteilung. Ebenso erzeugen die Verteilungen mit wenigen Ausprägungen geringfügig kleinere Spannweiten als die mit vielen Ausprägungen. Für $m = 6$ ergibt sich ein durchwachsenes Bild.

Ergänzend dazu zeigt die folgende Tabelle die empirischen Verteilungsfunktionen der Ränge bezüglich der Verteilungsabweichung der betrachteten Verfahren für $n = 100$.

$F(x)$	forest	mode	HD	kNN	irmi	MCA	EM_det	EM_stoch	EM_cat
≤1	0.358	0	0.108	0.244	0.017	0.014	0.122	0.139	0
≤2	0.5	0	0.278	0.556	0.092	0.025	0.278	0.281	0
≤3	0.692	0	0.431	0.794	0.233	0.064	0.403	0.372	0.011
≤4	0.797	0.003	0.578	0.936	0.458	0.156	0.528	0.539	0.014
≤5	0.917	0.014	0.708	0.969	0.728	0.247	0.717	0.692	0.019
≤6	0.964	0.1	0.847	0.994	0.978	0.431	0.831	0.836	0.025
≤7	0.969	0.144	0.994	1	1	0.981	0.939	0.939	0.042
≤8	1	0.744	1	1	1	1	0.983	1	0.272
≤9	1	1	1	1	1	1	1	1	1

Tabelle 7: Empirische Verteilungsfunktion der Ränge für VA und $n = 100$

Wie in der Tabelle ersichtlich ist, kann missForest am häufigsten die besten Ergebnisse erzielen. Dies ist in ca. 36 % aller Beobachtungen der Fall. Außerdem zählt es zu ca. 69 % zu den besten drei und zu 80 % zu den besten vier Verfahren, was die Konstanz der ver-

gleichsweise guten Ergebnisse unterstreicht. Am zweithäufigsten kann kNN die besten Ergebnisse erzielen. Sogar in ca. 79 % der Beobachtung ist kNN Teil der drei besten Verfahren und gehört damit auch fast immer zur Gruppe der besten. Sehr ähnlich ist die Verteilung der Ränge bei Random Hot-Deck und der deterministischen sowie stochastischen EM-Imputation. Alle drei Verfahren gehören mit knapp 40 % zu den besten drei und mit knapp 70 % zu den besten fünf Verfahren. Die Ergebnisse von irmi liegen überwiegend im Mittelfeld. Zu 72,8 % gehört es den besten fünf Verfahren an, wovon ca. 63,6 % auf die Ränge 3 bis 5 entfallen. MCA ist mit über 75 % Rang 6 oder schlechter. Die EM-Imputation speziell für kategoriale Daten und die Modusimputation erzielen mit Abstand am häufigsten die schlechtesten Ergebnisse.

Für $n = 500$ sind die Ergebnisse der Verteilungsabweichung in Abbildung 6 zu sehen. Ein Vergleich zwischen Abbildung 5 und Abbildung 6 verdeutlicht, dass auch bei der Verteilungsabweichung nur geringfügige Veränderungen mit der Erhöhung der Anzahl der Objekte n einhergehen. Insbesondere die deterministische und stochastische EM-Imputation können sich verbessern, vor allem in Fällen mit wenigen Ausprägungen bei einer ungleichen Verteilung unter dem MNAR-Ausfallmechanismus. Dennoch erzeugen sie hier erneut keine validen Ergebnisse für $p = 0{,}5$. Dies ist auch hier daran erkennbar, dass die Verteilungsabweichung hier größer als die verwendete Ausfallrate ist. Die Ergebnisse aller anderen Verfahren haben sich nur in einem geringen Maß verändert.

In Tabelle 8 sind erneut die durch die Erhöhung der Anzahl der Objekte bewirkten Verschlechterungen (negativ), Verbesserungen (positiv) und mittleren Änderungen angegeben.

	forest	mode	HD	kNN	irmi	MCA	EM_det	EM_stoch	EM_cat
min	-0,024	-0,054	-0,006	-0,024	-0,037	-0,063	-0,064	-0,047	-0,053
max	0,053	0,059	0,042	0,034	0,051	0,074	0,198	0,162	0,049
mean	0,008	0,001	0,011	0,01	0,009	0,004	0,002	0,01	0,003

Tabelle 8: Veränderung der Ergebnisse der VA durch die Erhöhung von n

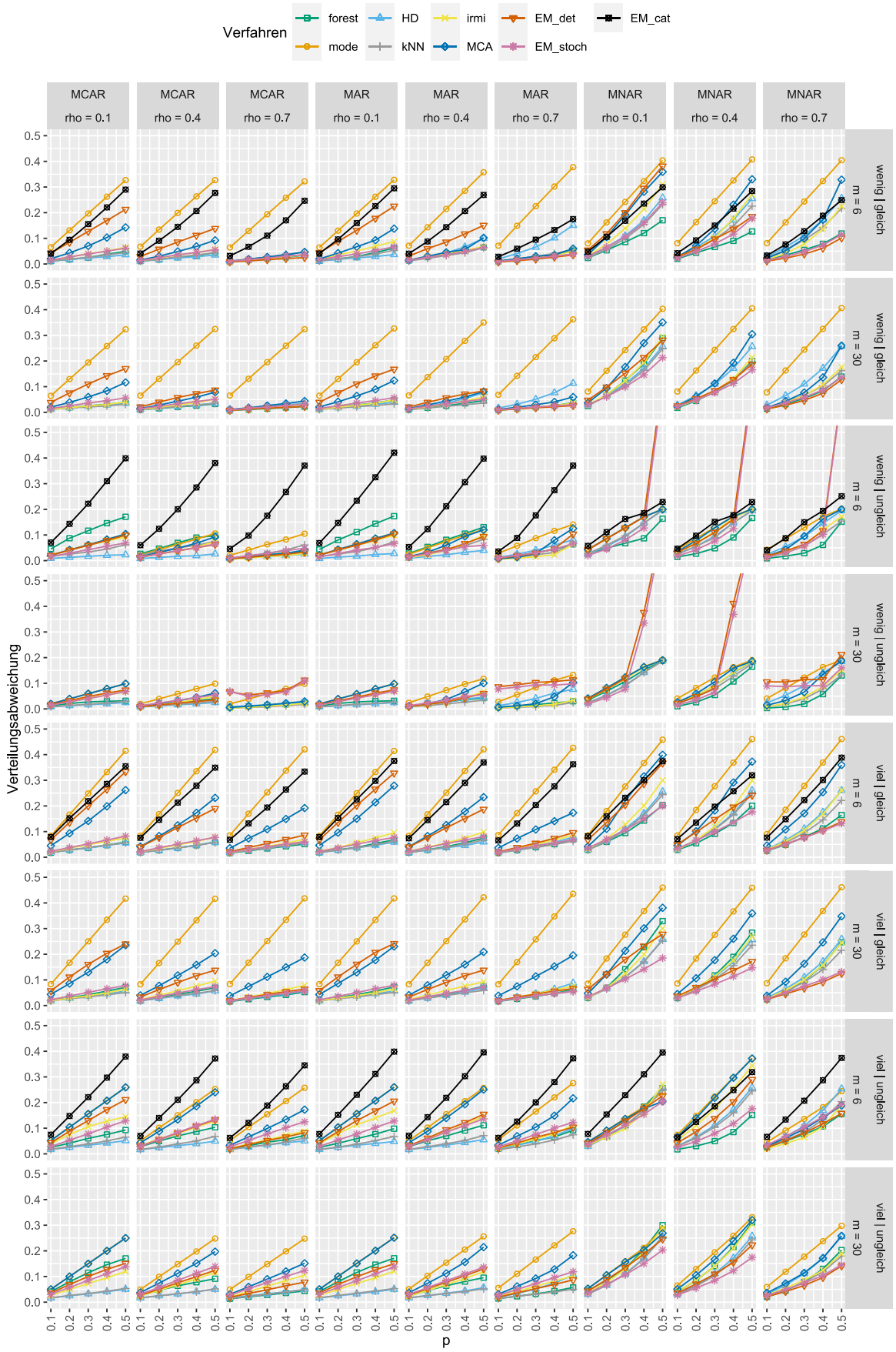

Abbildung 6: Verteilungsabweichung bei $n = 500$

Die größten Vorteile durch die Erhöhung der Objektanzahl zeigen sich bei missForest, Random Hot-Deck und irmi. Diese Verfahren erzielen für fast alle Faktorstufenkombinationen eine leichte Verbesserung. Hingegen zeigen die übrigen Verfahren kaum Veränderungen im Verhalten und ihre erzielten Ergebnisse bleiben sehr ähnlich. Eine Übersicht zur genauen Änderung der einzelnen Ränge für jedes Verfahren gibt die nachfolgende Tabelle 9.

f(*x*)	forest	mode	HD	kNN	irmi	MCA	EM_det	EM_stoch	EM_cat
1	-0,114	0	0,192	-0,025	0,019	-0,014	-0,022	-0,039	0
2	-0,019	0	-0,003	0,111	-0,003	-0,005	-0,106	0,019	0
3	0,133	0	-0,003	-0,066	0,037	-0,028	-0,058	0,006	-0,011
4	0,011	0	-0,036	-0,050	0,008	-0,048	0	0,111	-0,003
5	-0,059	-0,011	-0,052	0,014	0,122	-0,021	-0,009	0,014	-0,002
6	0,017	-0,014	-0,039	0,011	-0,172	0,135	0,103	-0,016	0,008
7	0,004	0,012	-0,056	0,005	-0,014	-0,033	0,089	-0,045	0,027
8	0,027	0,002	-0,003	0	0,003	0,014	0,012	-0,050	-0,016
9	0	0,011	0	0	0	0	-0,009	0	-0,003

Tabelle 9: Abweichung der relativen Häufigkeiten der Ränge für VA

Tabelle 9 zeigt, dass Random Hot Deck nun wesentlich häufiger den ersten Rang erreicht hat. Irmi schließt weniger häufig auf Rang 6 ab, zeigt jedoch eine vermehrte Platzierung auf Rang 5 sowie gelegentlich auf den Rängen 3 und 1. Obwohl missForest bessere Ergebnisse erzielt, belegt es nun gehäuft den dritten Rang und seltener die Ränge 1 und 2. Die stochastische EM-Imputation schneidet weniger häufig unter den letzten vier Rängen ab, belegt dafür jedoch vermehrt den vierten Rang. KNN profitiert nur geringfügig von der Erhöhung der Anzahl der Objekte, behält jedoch seine guten Ergebnisse bei und erreicht häufiger den zweiten Rang. Alle anderen Verfahren werden nun vermehrt im hinteren Rangbereich platziert.

Zusammenfassend hat die Anzahl der Objekte auch auf die Verteilungsabweichung nur einen geringen Einfluss, wobei die Verfahren davon mehr oder weniger stark betroffen sein können. Die Merkmalsanzahl wirkt sich unterschiedlich auf die Verfahren aus, so führt eine Erhöhung der Anzahl der Merkmale bei missForest zu einer Verschlechterung für den MNAR-Ausfallmechanismus, aber zu einer Verbesserung, wenn wenige Ausprägungen

und eine ungleiche Verteilung vorliegen. Auf die Modusimputation und Random Hot-Deck besitzt die Merkmalsanzahl beispielsweise nur eine sehr geringe Auswirkung. Alle weiteren Verfahren können unter bestimmten Umständen eher von der Erhöhung der Merkmalsanzahl profitieren, abgesehen von der EM-Imputation speziell für kategoriale Daten, wobei auch hier keine Ergebnisse erhoben werden können. Eine Verschlechterung der Ergebnisse wird wiederum durch die Erhöhung des Anteils der fehlenden Werte p bewirkt. Die Auswirkung auf die einzelnen Verfahren ist dabei sehr unterschiedlich. Verfahren wie missForest, Random Hot-Deck und kNN sind größtenteils robust gegen die Veränderungen, wobei die EM-Imputation für kategoriale Daten, MCA und die Modusimputation deutlich stärker von der Verschlechterung betroffen sind. Auch zwischen den verschiedenen Verteilungsformen sind Unterschiede erkennbar. So sind die Veränderungen bei wenigen Ausprägungen und einer ungleichen Verteilung verhältnismäßig gering (abgesehen von der EM-Imputation für kategoriale Daten). Insbesondere für den MNAR-Ausfallmechanismus sind die Verschlechterungen bzgl. des Anteils der fehlenden Werte p größer. Die Erhöhung des Zusammenhangs (ρ) besitzt einen positiven Effekt auf die Ergebnisse fast aller Verfahren. Dieser ist jedoch in der Regel geringer als für den PFC. Besonders im Fall des MNAR-Ausfallmechanismus kann die Erhöhung des Zusammenhangs eine Verbesserung der Ergebnisse bewirken. Beispiele dafür sind missForest, kNN und irmi. Weitestgehend unberührt bleiben von der Erhöhung des Zusammenhangs die Ergebnisse für die Modusimputation und Random Hot-Deck. Die Ergebnisse unter dem MCAR- und MAR-Ausfallmechanismus sind sich auch hier sehr ähnlich, lediglich Random Hot-Deck liefert bessere Ergebnisse unter dem MCAR-Ausfallmechanismus, wenn $\rho = 0,7$ ist. Fast alle Verfahren liefern unter dem MCAR- und dem MAR- bessere Ergebnisse als unter dem MNAR- Ausfallmechanismus, wobei die Unterschiede geringer sind als für den PFC. Die EM-Imputationsverfahren können nur für den Fall einer ungleichen Verteilung mit wenigen Ausprägungen bessere Ergebnisse unter dem MCAR- bzw. MAR-Ausfallmechanismus liefern. Hinsichtlich der Verteilungsformen werden in der Regel die besten Ergebnisse bei einer ungleichen Verteilung mit wenigen Ausprägungen erzielt, gefolgt von der ausgeglichenen Verteilung mit wenigen Ausprägungen, der unausgeglichenen Verteilung mit vie-

len Ausprägungen sowie der ausgeglichenen Verteilung mit vielen Ausprägungen. Von dieser Reihenfolge gibt es bei der Hälfte der Verfahren kleinere Abweichungen. So ist für missForest die Anzahl der Merkmale entscheidend. Für $m = 30$ trifft hier die Reihenfolge zu. Bei $m = 6$ hingegen ist die Reihenfolge wiederum abhängig vom Zusammenhang zwischen den Merkmalen. Hier kann die unausgeglichene Verteilung mit wenigen Ausprägungen am stärksten von einem steigenden Zusammenhang profitieren. Für kNN und irmi werden die Ergebnisse der verschiedenen Verteilungen sehr ähnlich, wenn die Anzahl der Objekte hoch ist. Die Verteilungen verhalten sich sowohl für die deterministische als auch die stochastische EM-Imputation sehr ähnlich. Hier fällt besonders auf, dass die Ergebnisse für die unausgeglichene Verteilung mit wenigen Ausprägungen unter MNAR besonders schlecht und zum Teil nicht valide sind. Für die EM-Imputation speziell für kategoriale Daten gilt die Reihenfolge auch nicht, da die ausgeglichene Verteilung mit wenigen Ausprägungen die besten Ergebnisse liefert.

3.3 Evaluation der Korrelation

Die Abbildung 7 und die Abbildung 8 zeigen die Auswirkungen der verschiedenen Faktorstufen auf die Korrelationsschätzung, welche in Form der RMSE des Zusammenhangs (Cramer's V) zwischen den Merkmalen vor der Imputation und nach der Imputation berechnet werden. Demzufolge ist ein Verfahren umso besser in der Wiederherstellung des ursprünglichen Zusammenhangs, je näher die Werte an Null liegen. Wie bei den vorherigen Gütekriterien werden nachfolgend die Ergebnisse zunächst für $n = 100$ gezeigt und anschließend für $n = 500$. Abbildung 7 zeigt die Ergebnisse für die Korrelationsschätzungen.

Generell hat die Erhöhung der Ausfallrate einen negativen Einfluss auf die Korrelationsschätzungen, wobei die Stärke dieses Einflusses von der spezifischen Verteilungsform abhängt. Zudem zeigt sich, dass die Wiederherstellung höherer Korrelationen im Vergleich zu niedrigeren Korrelationen schwieriger ist. Insbesondere im Fall weniger Ausprägungen erzielen die Verfahren unter dem MNAR-Ausfallmechanismus schlechtere Ergebnisse im Vergleich zu MCAR und MAR. Zusätzlich werden für Verteilungen mit wenigen Ausprägungen schlechtere Ergebnisse erzielt als für solche mit vielen Ausprägungen. Die spezifi-

schen Eigenheiten bezüglich einzelner Verfahren werden im Folgenden näher betrachtet. Die Modusimputation erzielt für $\rho = 0{,}1$ gute Ergebnisse, allerdings verschlechtern sich diese mit zunehmender Korrelation. Auch kNN kann überwiegend gute Ergebnisse erzielen. Im Gegensatz zur Modusimputation sind diese jedoch verhältnismäßig schlecht, wenn $\rho = 0{,}1$. Besser performt kNN, wenn der ursprüngliche Zusammenhang stärker ist. Eine Ausnahme bildet die ungleiche Verteilung mit vielen Ausprägungen und $m = 6$, bei der die Ergebnisse im Vergleich zu den anderen Verfahren generell schlecht ausfallen. MCA erzielt insbesondere gute Ergebnisse, wenn die Verteilung unausgeglichen ist. Bei einer ausgeglichenen Verteilung hingegen verbessern sich die Ergebnisse mit zunehmendem Zusammenhang. Irmi gehört häufig zu den besten Verfahren. Bei einer ausgeglichenen Verteilung verbessern sich die Ergebnisse mit zunehmendem ursprünglichem Zusammenhang. Zudem wirkt sich eine Erhöhung der Merkmalsanzahl positiv auf das Verfahren aus. Durchwachsene Ergebnisse erzielen die deterministische und die stochastische EM-Imputation. Besonders schlecht sind diese für wenige Ausprägungen und $\rho = 0{,}7$. missForest, Random Hot Deck und die EM-Imputation speziell für kategoriale Daten können nur schwierig die ursprüngliche Korrelation wiederherstellen. Insbesondere missForest erzielt schlechte Ergebnisse, wenn viele Ausprägungen vorhanden sind. Ebenso wie die Modusimputation ist Random Hot-Deck kaum dazu in der Lage den ursprünglichen Zusammenhang wiederherzustellen, wenn $\rho > 0{,}1$ ist. Gleiches gilt für die EM-Imputation speziell für kategoriale Daten, die mit zunehmendem Zusammenhang schlechtere Ergebnisse liefert.

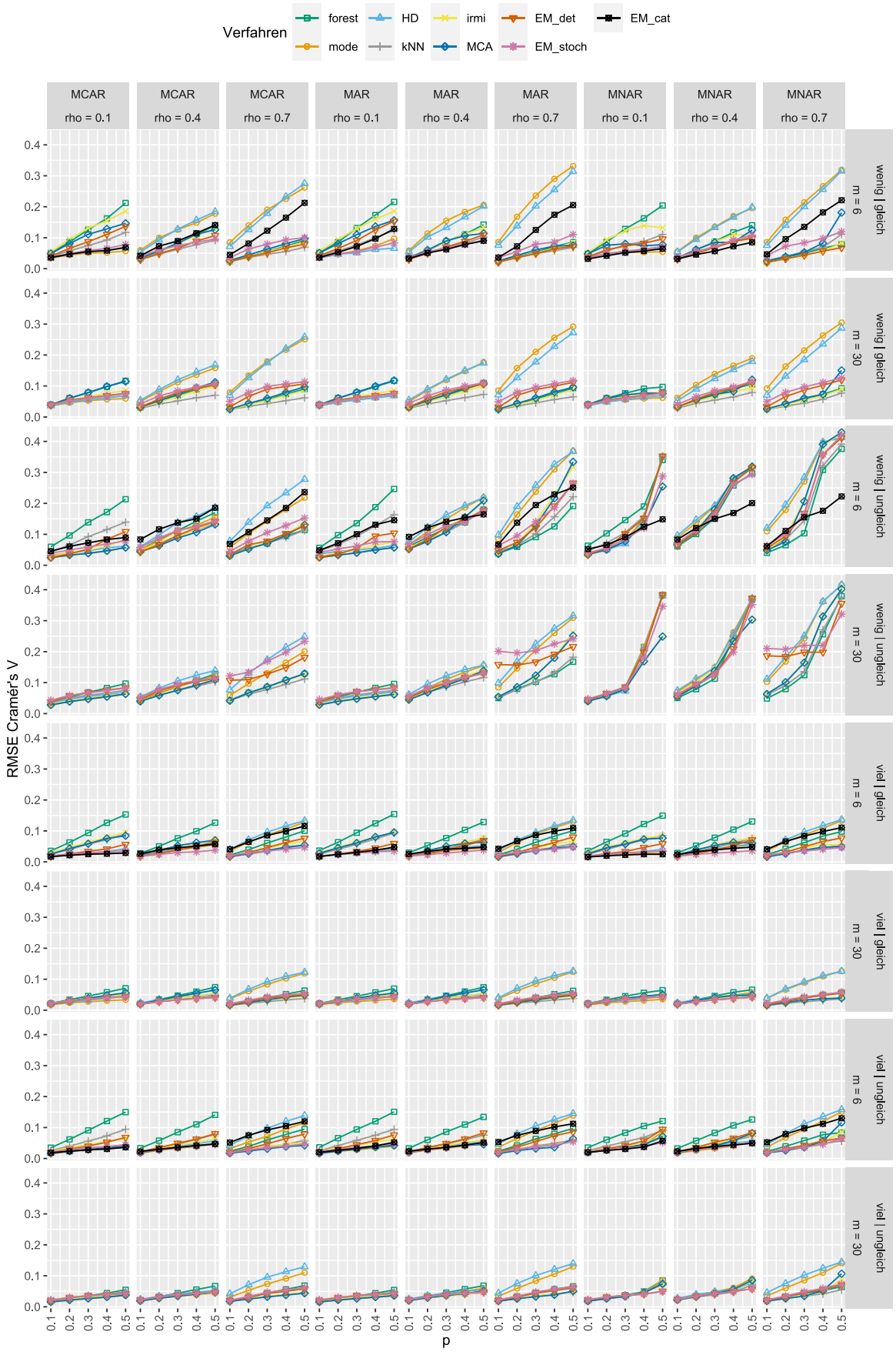

Abbildung 7: RMSE von Cramér's V bei $n = 100$

Die nachfolgende Tabelle fasst die maximalen Spannweiten und Mittelwerte der besten Verfahren aller Teilabbildung aus Abbildung 7 zusammen.

	MCAR	MCAR	MCAR	MAR	MAR	MAR	MNAR	MNAR	MNAR		
	rho = 0,1	rho = 0,4	rho = 0,7	rho = 0,1	rho = 0,4	rho = 0,7	rho = 0,1	rho = 0,4	rho = 0,7		
max SP	0,1554	0,0921	0,2059	0,1497	0,1156	0,2628	0,1503	0,1135	0,2509	m = 6	w | g
best mean	0,0475	0,0616	0,046	0,0521	0,0618	0,0462	0,0465	0,0579	0,044		
max SP	0,0573	0,0977	0,1964	0,0493	0,1039	0,2266	0,035	0,1105	0,2277	m = 30	w | g
best mean	0,0501	0,0508	0,0432	0,0539	0,0517	0,0452	0,0524	0,0542	0,0477		
max SP	0,1561	0,0546	0,1651	0,1892	0,0556	0,2004	0,2041	0,1177	0,2218	m = 6	w | u
best mean	0,04	0,0867	0,0711	0,0411	0,1078	0,1005	0,0845	0,1363	0,1214		
max SP	0,0341	0,0364	0,137	0,0342	0,0402	0,1511	0,1354	0,0701	0,1637	m = 30	w | u
best mean	0,0458	0,0729	0,0776	0,0461	0,0836	0,1046	0,1171	0,1487	0,1544		
max SP	0,1246	0,0886	0,0865	0,1196	0,0903	0,0872	0,1248	0,0944	0,0902	m = 6	v | g
best mean	0,0236	0,0285	0,0328	0,0275	0,0283	0,0335	0,0214	0,0277	0,0329		
max SP	0,0374	0,0336	0,0851	0,0339	0,0335	0,0889	0,0292	0,0241	0,0894	m = 30	v | g
best mean	0,0266	0,0316	0,0273	0,0275	0,0315	0,0273	0,0269	0,0318	0,0266		
max SP	0,1137	0,0952	0,0965	0,111	0,0887	0,0908	0,0695	0,0768	0,1017	m = 6	v | u
best mean	0,0261	0,0322	0,0306	0,0282	0,0329	0,0328	0,0328	0,0331	0,0364		
max SP	0,0177	0,0238	0,0855	0,0185	0,0223	0,0924	0,0365	0,0336	0,0898	m = 30	v | u
best mean	0,0263	0,0323	0,0312	0,026	0,034	0,0325	0,0334	0,0397	0,0353		

Tabelle 10: Maximale Spannweite und Mittelwerte für Cramér's V und $n = 100$

Es fällt auf, dass die Spannweite für $\rho = 0,3$ in den meisten Fällen am geringsten ist, gefolgt von $\rho = 0,1$ und $\rho = 0,7$. Ebenso verhält es sich für die Mittelwerte der besten Ergebnisse, wobei diese für den Fall von $\rho = 0,3$ am höchsten sind. Die maximalen Spannweiten verhalten sich über alle Ausfallmechanismen hinweg sehr ähnlich. Gleiches gilt wiederum für die Mittelwerte der besten Ergebnisse, wobei die Mittelwerte für die ungleiche Verteilung mit wenigen Ausprägungen für den MNAR-Ausfallmechanismus deutlich höher sind als unter MCAR und MAR. Die maximalen Spannweiten sind insbesondere für $\rho = 0,1$ deutlich geringer für $m = 30$ als für $m = 6$. Hinsichtlich der Mittelwerte sind diese abgesehen von vereinzelten Ausnahmen sehr ähnlich für $m = 6$ und $m = 30$. Die Art der Verteilung besitzt auch auf die Korrelationsschätzung einen erheblichen Einfluss. Sowohl die Spannweiten als auch die Mittelwerte unterscheiden sich zwischen den Verteilungsarten zum Teil deutlich. Dabei sind die maximalen Spannweiten und die Mittelwerte der besten Verfahren geringer für Verteilungen mit vielen Ausprägungen. Des Weiteren zeigen sich überwiegend geringere maximale Spannweiten bei der ungleichen Verteilung im Ver-

gleich zur ausgeglichenen Verteilung, wobei jedoch die Mittelwerte der letzteren größtenteils besser sind.

Die nachfolgende Tabelle zeigt die empirischen Verteilungsfunktionen der Ränge bezüglich der Wiederherstellung der ursprünglichen Korrelation durch die einzelnen Verfahren für $n = 100$.

$F(x)$	forest	mode	HD	kNN	irmi	MCA	EM_det	EM_stoch	EM_cat
≤1	0.056	0.264	0.028	0.253	0.036	0.15	0.056	0.094	0.089
≤2	0.086	0.339	0.114	0.383	0.217	0.353	0.167	0.219	0.128
≤3	0.203	0.383	0.194	0.536	0.422	0.483	0.275	0.325	0.178
≤4	0.297	0.422	0.328	0.622	0.578	0.617	0.447	0.464	0.225
≤5	0.406	0.467	0.406	0.744	0.675	0.731	0.744	0.586	0.242
≤6	0.556	0.517	0.461	0.844	0.836	0.794	0.858	0.875	0.258
≤7	0.611	0.758	0.553	0.919	0.897	0.964	0.961	0.944	0.392
≤8	0.772	0.928	0.858	1	0.981	0.994	0.997	1	0.472
≤9	1	1	1	1	1	1	1	1	1

Tabelle 11: Empirische Verteilungsfunktion der Ränge für Cramer's V und $n = 100$

Die Modusimputation kann am häufigsten die besten Ergebnisse erzielen. Dies ist in 26,4 % aller Faktorstufenkombinationen der Fall. In fast 50 % der Fälle ist es jedoch auf den Rängen 7, 8 oder 9. Dies liegt an der Tatsache, dass die Modusimputation für den Fall eines geringen Zusammenhangs ($\rho = 0{,}1$), was auf einen Anteil von 1/3 aller Ergebnisse zutrifft, gute Werte liefert, jedoch kaum in der Lage dazu ist, stärkere Zusammenhänge wiederherzustellen. Am zweithäufigsten kann kNN den ursprünglichen Zusammenhang am besten wiederherstellen. In 25,3 % alle Faktorstufenkombinationen ist das der Fall. Die soliden Ergebnisse von kNN werden dadurch gestützt, dass es in 62,2 % der Fälle zu den besten vier Verfahren gehört. MCA kann am dritthäufigsten die besten Ergebnisse liefern. Es ist außerdem in 48,3 % der Fälle unter den besten drei und in 61,7 % der Fälle unter den besten vier Verfahren. Relativ selten das beste Verfahren, aber dennoch häufig Teil der besten Verfahren, ist irmi. Es ist zu 42,2 % eines der drei besten Verfahren und zu 57,8 % unter den vier besten. Die deterministische und die stochastische EM-Imputation zählen zu über 40 % zu den besten vier Verfahren, wobei die deterministische EM-Imputation deutlich häufiger zu den besten fünf Verfahren zählt als das stochastische Pendant. Weniger

gute Ergebnisse erzeugen dagegen missForest, Random Hot-Deck und die EM-Imputation speziell für kategoriale Daten. Alle drei Verfahren zählen überwiegend zu den schlechteren Verfahren.

Für $n = 500$ sind die Ergebnisse in Abbildung 8 zu sehen. Ein Vergleich der Abbildung 7 mit der Abbildung 8 zeigt, dass die Ergebnisse für den Zusammenhang mit $n = 500$ überwiegend besser sind als die für $n = 100$, aber dennoch sehr ähnlich. Bei genauerer Betrachtung der Abbildung 8 fällt auf, dass die Modusimputation und Random Hot-Deck weiterhin mit Abstand die schlechtesten Ergebnisse für $\rho > 0{,}1$ erzeugen. Für die unausgeglichene Verteilung mit wenigen Ausprägungen unter dem MNAR-Ausfallmechanismus zeigt die EM-Imputation speziell für kategoriale Daten nun deutlich verbesserte Ergebnisse für $\rho = 0{,}7$. Das Verhalten aller anderen Verfahren hat sich nur geringfügig verändert.

Abgesehen vom Vergleich der Abbildung 7 und Abbildung 8 wird dies auch in Tabelle 12 deutlich. In ihr sind erneut die durch die Erhöhung der Anzahl der Objekte bewirkten Verschlechterungen (negative Werte), Verbesserungen (positive Werte) und mittleren Änderungen angegeben.

	forest	mode	HD	kNN	irmi	MCA	EM_det	EM_stoch	EM_cat
min	-0.026	-0.065	-0.077	-0.03	-0.041	-0.068	-0.058	-0.055	-0.062
max	0.157	0.093	0.091	0.104	0.122	0.085	0.137	0.113	0.127
mean	0.017	0.006	0.002	0.025	0.012	0.009	0.015	0.014	0.008

Tabelle 12: Veränderung der Ergebnisse von Cramer's V durch die Erhöhung von *n*

Im Allgemeinen zeigt kNN die stärkste Verbesserung, gefolgt von missForest, welches im Gesamten die zweitstärkste Verbesserung aufweisen kann. Die deterministische und stochastische EM-Imputation können an dritt- und viertstärksten von einer erhöhten Anzahl an Objekten profitieren. Ebenfalls zeigt sich bei irmi eine durchschnittliche Verbesserung der Ergebnisse. Die Veränderungen bei MCA und der EM-Imputation speziell für kategoriale Daten sind minimal, während die Modusimputation und Random Hot-Deck am wenigsten von der Anzahl der Objekte profitieren.

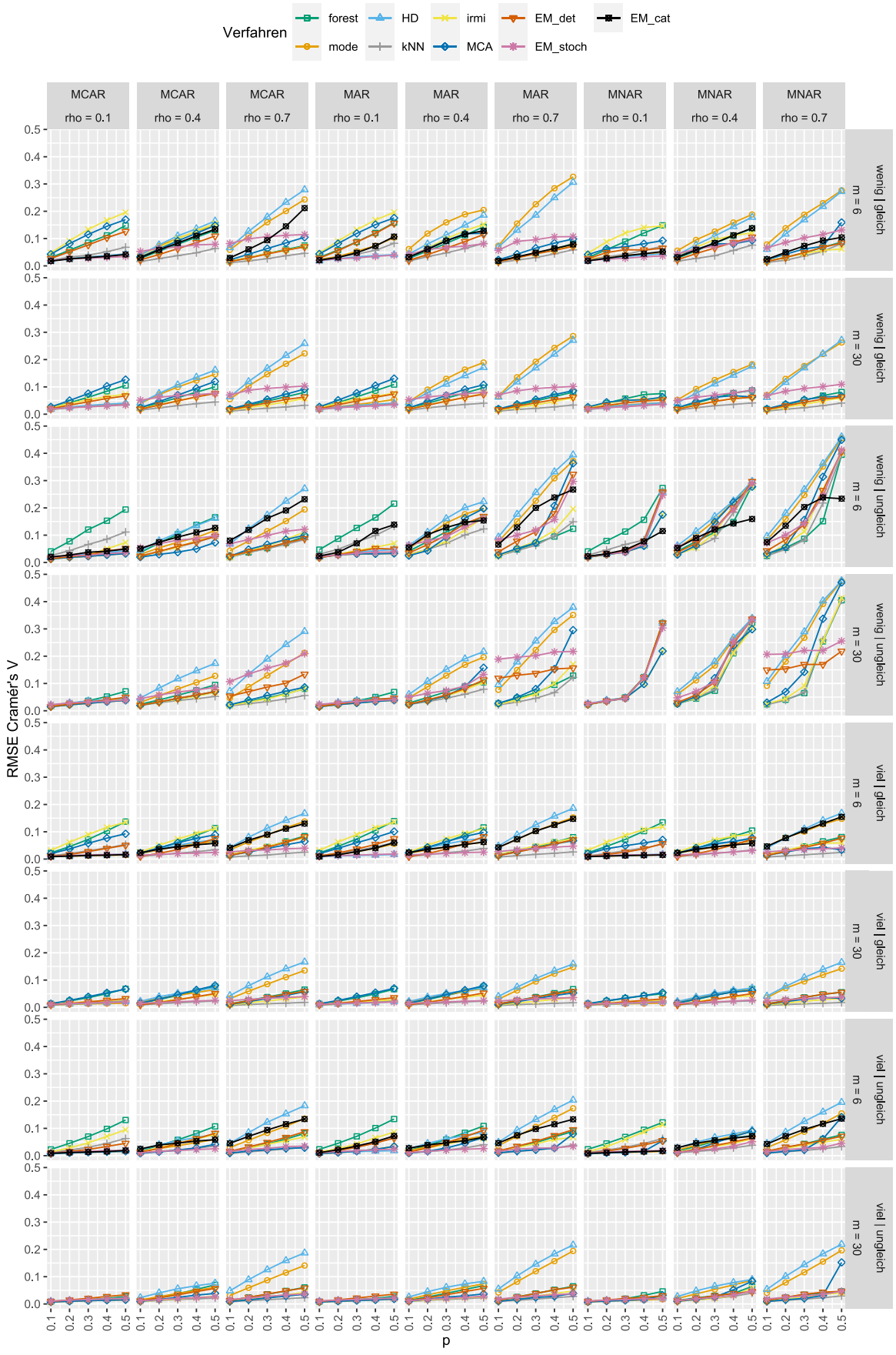

Abbildung 8: RMSE von Cramér's V bei $n = 500$

Tabelle 13 gibt einen Überblick über die empirische Verteilungsfunktion der Rangfolge eines jeden Verfahrens bezüglich Cramer's V.

f(x)	forest	mode	HD	kNN	irmi	MCA	EM_det	EM_stoch	EM_cat
1	-0,031	-0,122	0,011	0,278	-0,008	-0,031	-0,045	0,020	-0,067
2	0,012	-0,003	-0,028	0,011	0,022	-0,044	0,011	0,022	-0,003
3	-0,037	-0,002	0,001	-0,100	0,031	-0,011	0,048	0,064	0,009
4	0,051	0,044	-0,068	-0,030	-0,015	0,002	0,022	-0,017	0,005
5	0,096	0,024	-0,028	-0,070	-0,016	0	-0,033	-0,028	0,053
6	-0,019	0,023	0,007	-0,008	-0,067	0,043	0,072	-0,078	0,028
7	0,100	0,070	0,030	-0,036	-0,016	-0,051	-0,069	0,001	-0,025
8	-0,066	-0,020	0,028	-0,045	-0,018	0,098	-0,003	0,005	0,014
9	-0,106	-0,014	0,047	0	0,087	-0,006	-0,003	0,011	-0,014

Tabelle 13: Abweichung der relativen Häufigkeiten der Ränge für Cramer's V

Die Verbesserungen haben auch deutliche Auswirkungen auf die relativen Ranghäufigkeiten von kNN. Es ist 27,8 Prozentpunkte häufiger auf Rang 1 als zuvor. Trotz der Verbesserung gehört missForest weiterhin überwiegend zur Gruppe der schlechteren Verfahren, was sich auch anhand der Veränderung der relativen Ranghäufigkeiten zeigt. Die deterministische und die stochastische EM-Imputation werden nun häufiger in den Rängen zwei und drei platziert und sind somit vermehrt im oberen Mittelfeld der Verfahren vertreten. MCA befindet sich nun weniger unter den besten drei und dafür auch häufiger im Mittelfeld der Verfahren. Irmi konnte sich durchschnittlich verbessern und ist damit öfter auf den Rängen zwei und drei vertreten Nach der Erhöhung der Objektanzahl sind die EM-Imputation für kategoriale Daten und die Modusimputation häufiger im Mittelfeld platziert, während Random Hot Deck nun vermehrt im unteren Rangbereich abschließt.

Zusammenfassend führt die Erhöhung der Anzahl der Objekte zu einer leichten Verbesserung der Ergebnisse. Insbesondere, wenn die Verteilung unausgeglichen ist und nur wenige Objekte vorliegen, ist dieser Effekt am stärksten. Von der Erhöhung der Merkmalsanzahl können missForest und irmi am stärksten profitieren. Bei den restlichen Verfahren liegen nur eine geringe Verbesserung oder durchwachsene Ergebnisse vor. Lediglich bei der stochastischen EM-Imputation verschlechtern sich die Ergebnisse bei wenigen Ausprägungen und einer unausgeglichenen Verteilung. Die Ergebnisse aller Verfahren verschlechtern

sich mit zunehmendem Anteil fehlender Werte, jedoch werden die Verfahren unterschiedlich stark und zudem abhängig von den anderen Faktoren beeinflusst. Weniger stark werden die Verfahren vom Anteil der fehlenden Daten beeinflusst, wenn viele Ausprägungen vorliegen. Auch die Verfahren selbst reagieren unterschiedlich stark auf die Veränderung des Anteils der fehlenden Werte. So reagieren die Modusimputation und Random Hot-Deck am stärksten auf die Erhöhung des Anteils der fehlenden Werte, wenn der Zusammenhang $\rho > 0{,}1$ ist. Im Gegensatz zu den zuvor betrachteten Gütekriterien wird mit dem RMSE von Cramer's V betrachtet, inwiefern die Verfahren dazu in der Lage sind, verschiedene Stärken des Zusammenhangs wiederherzustellen. Dieser Sachverhalt erklärt, warum bei einigen Verfahren mit steigendem Zusammenhang die Ergebnisse bzgl. Cramer'S V schlechter werden. Dies betrifft die Modusimputation, Random Hot-Deck und die EM-Imputation speziell für kategoriale Daten. Für irmi, MCA und die beiden verbliebenen EM-Imputationsverfahren trifft dieser Effekt nur im Fall einer unausgeglichenen Verteilung mit wenigen Ausprägungen zu. Für missForest verbessern sich die Ergebnisse mit zunehmendem Zusammenhang, wenn $m = 6$ ist und auf kNN haben die unterschiedlichen Zusammenhänge kaum einen Effekt. Lediglich unter dem MNAR-Ausfallmechanismus und einer unausgeglichenen Verteilung mit wenigen Ausprägungen kann kNN geringe Zusammenhänge besser wiederherstellen als mittlere oder starke. Zwischen den Ergebnissen unter dem MCAR- und MAR-Ausfallmechanismus existieren kaum Unterschiede. Lediglich wenn die Verteilung unausgeglichen ist und nur wenige Ausprägungen vorliegen, liefern fast alle Verfahren bessere Ergebnisse unter dem MCAR-Ausfallmechanismus. Ebenso verhält es sich beim Vergleich von MAR- mit dem MNAR-Ausfallmechanismus, wobei hier die besseren Ergebnisse für den MAR-Ausfallmechanismus vorliegen. In Bezug auf die Verteilungsform können alle Verfahren die besten Ergebnisse erzielen, wenn viele Ausprägungen vorliegen. Besonders schlecht werden die Ergebnisse, wenn wenige Ausprägungen und eine unausgeglichene Verteilung gegeben sind sowie der Ausfallmechanismus MNAR vorliegt. Irmi, MCA sowie die deterministische und die stochastische EM-Imputation liefern außerdem auffallend schlechte Ergebnisse für die unausgeglichene Verteilung mit wenigen Ausprägungen, wenn der Ausfallmechanismus MAR und $\rho = 0{,}7$ ist.

Für irmi verschlechtern sich die Ergebnisse außerdem, wenn eine ausgeglichene Verteilung sowie $\rho = 0{,}1$ bzw. $\rho = 0{,}4$ vorliegen.

3.4 Evaluation der Regressionskoeffizienten

Im Gegensatz zu den bisher betrachteten Gütekriterien werden hinsichtlich der Auswirkungen auf die logistische Regression nur die Ergebnisse für $n = 500$ betrachtet. In Abbildung 9 sind die Ergebnisse in Form der RMSE zwischen den Regressionskoeffizienten vor der Imputation und nach der Imputation abgebildet. Somit spricht ein niedriger RMSE für eine bessere Wiederherstellung der ursprünglichen Ergebnisse. Eine Besonderheit dieser Abbildung ist, dass einzelne Datenpunkte nicht dargestellt werden. Dies ist der Fall, wenn eine Regression aufgrund der durch die Imputation erzeugten Multikollinearität oder zu wenigen Beobachtungen in bestimmten Ausprägungen nicht durchgeführt werden konnte, bzw. gescheitert ist. Es müssen 210 der 3240 Datenpunkte entfernt werden, abzüglich der 180 entfallenen Datenpunkte für die EM-Imputation für kategoriale Daten entspricht dies ca. 0,96 %. Das Fehlen dieser Datenpunkte führt nicht zwangsläufig zu einem Informationsverlust, da die angewandten Verfahren keine sinnvollen Ergebnisse hinsichtlich des betrachteten Gütekriteriums erzeugen können.

In Abbildung 9 ist erkennbar, dass sich die Ergebnisse aller Verfahren sehr stark ähneln. Lediglich für die ungleiche Verteilung mit wenigen Ausprägungen sind deutliche Unterschiede zu sehen. Insbesondere unter dem MNAR-Ausfallmechanismus sind auffallend schlechte Werte erkennbar. Darüber hinaus werden bei dieser Verteilungsform und $m = 30$ schlechte Ergebnisse erzeugt, wenn $\rho = 0{,}7$ ist. Dies tritt auch unter den MAR- und MCAR-Ausfallmechanismen auf, wenn $m = 6$ ist.

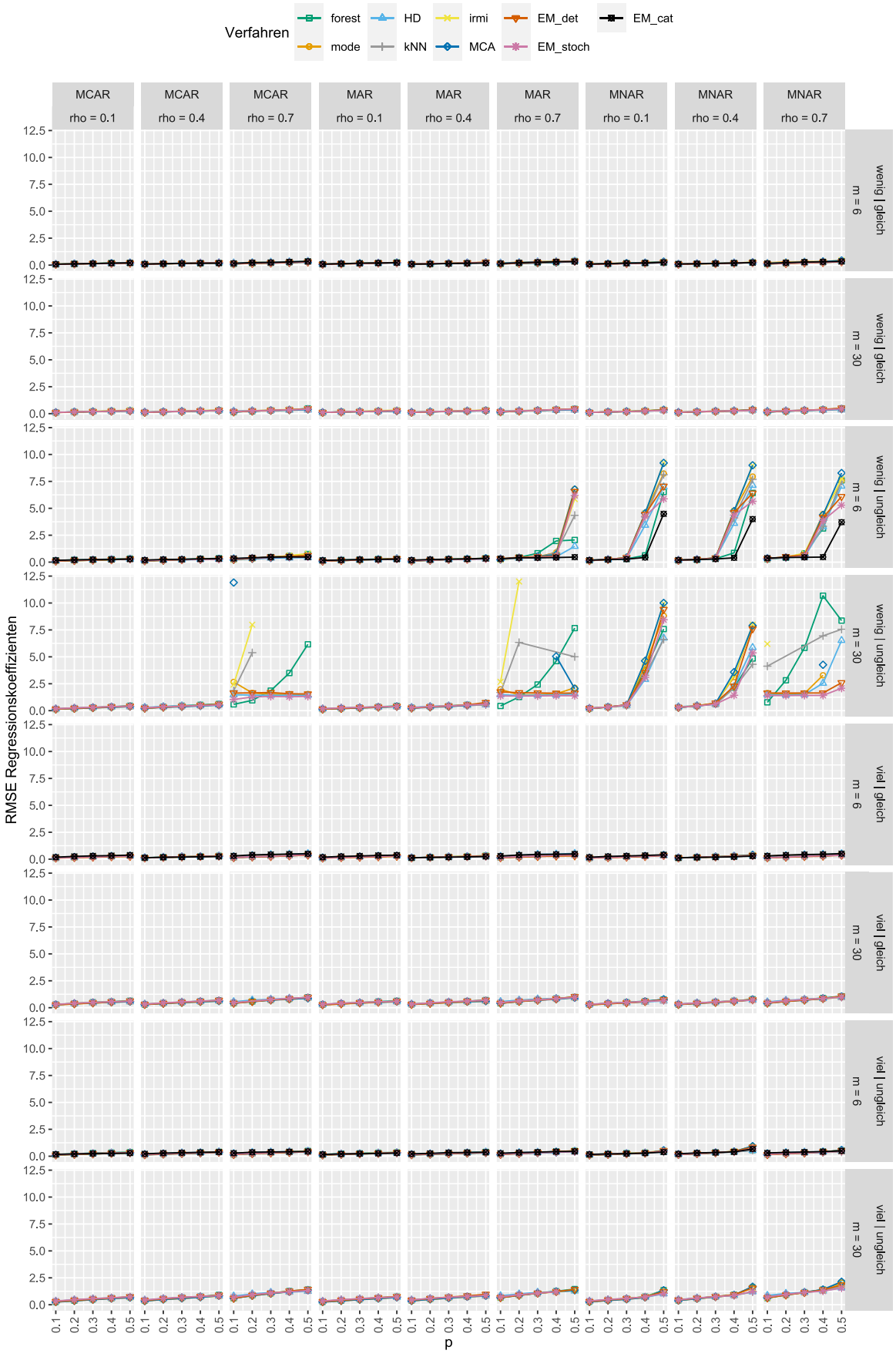

Abbildung 9: RMSE zwischen den Regressionskoeffizienten bei $n = 500$

Für eine bessere Vergleichbarkeit der Ergebnisse werden in der folgenden Tabelle die maximalen Spannweiten und die Mittelwerte der besten Ergebnisse aller Teilabbildung aus Abbildung 9 aufgeführt.

	MCAR	MCAR	MCAR	MAR	MAR	MAR	MNAR	MNAR	MNAR		
	rho = 0,1	rho = 0,4	rho = 0,7	rho = 0,1	rho = 0,4	rho = 0,7	rho = 0,1	rho = 0,4	rho = 0,7		
max SP	0,0578	0,0337	0,0983	0,0397	0,0834	0,1246	0,1084	0,0425	0,1459	m = 6	w | g
best mean	0,1208	0,1358	0,1922	0,1389	0,1354	0,205	0,1382	0,1432	0,1956		
max SP	0,0748	0,0465	0,1482	0,0861	0,0558	0,1108	0,074	0,0622	0,1572	m = 30	w | g
best mean	0,1741	0,1995	0,2614	0,1739	0,1944	0,2628	0,1922	0,207	0,2687		
max SP	0,0752	0,0872	0,3696	0,0748	0,0743	6,3272	4,7302	4,9997	4,5692	m = 6	w | u
best mean	0,1814	0,205	0,346	0,1876	0,2158	0,376	1,1147	1,0133	1,0636		
max SP	0,0843	0,1126	11,3164	0,0641	0,2201	10,725	3,3818	3,6794	9,2645	m = 30	w | u
best mean	0,2536	0,3687	1,0947	0,2514	0,3881	1,1693	2,1049	1,4125	1,4176		
max SP	0,1255	0,0612	0,1959	0,1352	0,0751	0,2258	0,1213	0,1167	0,1963	m = 6	v | g
best mean	0,181	0,1891	0,2537	0,1794	0,1849	0,2303	0,2003	0,1941	0,2394		
max SP	0,0882	0,0719	0,1449	0,0795	0,1084	0,1593	0,1512	0,0877	0,124	m = 30	v | g
best mean	0,4187	0,4618	0,6661	0,4167	0,4574	0,6749	0,4466	0,4937	0,6939		
max SP	0,0893	0,1098	0,1376	0,0765	0,1076	0,1088	0,1677	0,4018	0,1944	m = 6	v | u
best mean	0,2054	0,2394	0,2874	0,2063	0,221	0,2985	0,2274	0,3256	0,2828		
max SP	0,1065	0,1131	0,2084	0,115	0,1661	0,1852	0,3469	0,4938	0,5898	m = 30	v | u
best mean	0,4645	0,5817	0,9771	0,467	0,5948	1,0117	0,5713	0,7421	1,0902		

Tabelle 14: Maximale Spannweite und Mittelwerte für Cramér's V und $n = 100$

Die kleinsten maximalen Spannweiten treten bei $\rho = 0,4$ auf, gefolgt von $\rho = 0,1$ und $\rho = 0,7$. In Bezug auf die Mittelwerte der besten Ergebnisse sind diese bei $\rho = 0,1$ am niedrigsten, gefolgt von $\rho = 0,4$ und $\rho = 0,7$. Die Spannweiten und Mittelwerte ähneln sich weitgehend bezogen auf die Ausfallmechanismen. Eine Ausnahme bildet die unausgeglichene Verteilung mit wenigen Ausprägungen, bei der sowohl die Spannweiten als auch die Mittelwerte deutlich unter dem MNAR-Ausfallmechanismus ansteigen. Außerdem weisen die Verteilungen mit einer geringeren Anzahl von Merkmalen in der Regel eine geringere maximale Spannweite sowie niedrigere Mittelwerte der besten Ergebnisse auf.[13] Bezüglich der maximalen Spannweite und der Mittelwerte der besten Verfahren weisen die Verteilungen in der folgenden Reihenfolge die niedrigsten Werte auf: unausgeglichen mit wenigen Ausprägungen, unausgeglichen mit vielen Ausprägungen, ausgeglichen mit wenigen Ausprägungen und ausgeglichen mit vielen Ausprägungen.

[13] Dies ist nicht überraschend, da aufgrund der geringeren Anzahl von Merkmalen auch eine deutlich geringere Anzahl von Koeffizienten geschätzt werden musste.

Die empirische Verteilungsfunktion der Ränge aller Verfahren sind in Tabelle 15 zu sehen. Alle Verfahren, die keine sinnvollen Ergebnisse liefern können, werden bereinigt. Diese bereinigten Werte stellen bei der Rangvergabe für die entsprechende Faktorstufenkombination den Rang 9 dar. Zusätzlich ist der Anteil der bereinigten Werte (NA) in Tabelle 15 zu sehen.

$F(x)$	forest	mode	HD	kNN	irmi	MCA	EM_det	EM_stoch	EM_cat
≤1	0,039	0,183	0,083	0,056	0,017	0,206	0,247	0,067	0,081
≤2	0,111	0,4	0,172	0,117	0,056	0,511	0,369	0,164	0,094
≤3	0,272	0,469	0,297	0,217	0,167	0,631	0,525	0,303	0,114
≤4	0,408	0,558	0,375	0,369	0,369	0,742	0,669	0,375	0,131
≤5	0,511	0,617	0,469	0,642	0,539	0,786	0,761	0,519	0,156
≤6	0,614	0,689	0,569	0,839	0,664	0,836	0,847	0,728	0,189
≤7	0,778	0,833	0,692	0,95	0,789	0,872	0,928	0,861	0,253
≤8	0,928	0,914	0,944	0,975	0,914	0,939	0,986	0,994	0,317
≤9	1	1	1	1	1	1	1	1	1
NA	0	0.003	0	0.019	0.028	0.031	0	0.003	0.5

Tabelle 15: Empirische Verteilungsfunktion der Ränge für die Regression und $n = 100$

Am häufigsten zu den besten bzw. besten drei Verfahren zählen MCA, die Modusimputation und die deterministische EM-Imputation. Alle weiteren Verfahren bis auf die EM-Imputation speziell für kategoriale Daten befinden sich überwiegend im mittleren Rangbereich. Diese schließt hauptsächlich auf den hinteren Rängen ab. Abgesehen davon sind unter den letzten drei Rängen verhältnismäßig oft missForest, Rand Hot Deck und irmi vertreten. Zu beachten ist dabei, dass, wie der Abbildung 9 bzw. Tabelle 14 zu entnehmen ist, die Abstände zwischen den Verfahren sehr gering sind und damit weniger Aussagekraft besitzen.

Zusammenfassend erzeugen alle Verfahren für die meisten Faktorkombinationen in Bezug auf die Regressionskoeffizienten für $n = 500$ sehr gute Ergebnisse. Jedoch zeigen sich deutliche Unterschiede bei der ungleichen Verteilung mit wenigen Ausprägungen. Unter dem MNAR-Ausfallmechanismus sind die Resultate deutlich schlechter. Für $m = 30$ und $\rho = 0,7$ treten diese ungünstigen Ergebnisse auch bei MCAR und MAR auf. Bei $m = 6$ sind unter derselben Korrelation ebenfalls schlechte Werte beim MAR-Ausfallmechanismus zu beobachten.

3.5 Verlässlichkeit der Ergebnisse

Zum Nachweis der Verlässlichkeit der Ergebnisse werden für alle Gütekriterien, für jedes Verfahren sowie für alle Faktorstufenkombinationen die Monte-Carlo-Standardfehler berechnet. Dabei werden für alle Gütekriterien die maximalen Monte-Carlo-Standardfehler, die Durchschnittswerte sowie der Median der Monte-Carlo-Standardfehler für jedes Verfahren angegeben. Zusätzlich ist die verfahrensspezifische Spannweite SP hinsichtlich des betrachteten Gütekriteriums angegeben. Dazu wird für jedes Verfahren der minimale Wert vom maximalen Wert des betrachteten Gütekriteriums subtrahiert. Darüber hinaus wird in Anlehnung an Rockel (2022, 184 f.) das Verhältnis zwischen dem maximalen Monte-Carlo-Standardfehler $\hat{\sigma}_{\mathrm{MC,max}}$ und der Spannweite SP eines Gütekriteriums für das jeweilige Verfahren gebildet sowie die Länge des Konfidenzintervalls für den Mittelwert mit dem größten Monte-Carlo-Standardfehler $l_{\max}$ angegeben. Diese Konfidenzintervalle können bei beliebig verteilten Grundgesamtheiten für den Erwartungswert eines Gütekriteriums für eine bestimmte Faktorstufenkombination eines Verfahrens in folgender Form bestimmt werden (vgl. Rockel, 2022, S. 185).

$$\left[\bar{\theta} - \hat{\sigma}_{\mathrm{MC}} \cdot z_{1-\frac{\alpha}{2}}; \bar{\theta} + \hat{\sigma}_{\mathrm{MC}} \cdot z_{1-\frac{\alpha}{2}}\right]$$

Dabei ist $\bar{\theta}$ der Mittelwert für ein Gütekriterium der jeweiligen Faktorstufenkombination eines Verfahrens über alle Wiederholungen, $\hat{\sigma}_{\mathrm{MC}}$ der dazugehörige Monte-Carlo-Standardfehler und $z_{1-\alpha/2}$ das $(1-\frac{\alpha}{2})$-Fraktil der Standardnormalverteilung. Dabei ergibt sich die maximale Länge des Konfidenzintervalls bei einer Sicherheit von 95 % mit $l_{\max} = 2 \cdot \hat{\sigma}_{\mathrm{MC,max}} \cdot z_{0,975}$. Mit den Quotienten aus den maximalen Monte-Carlo-Standardfehlern und den Spannweiten können erstere ins Verhältnis zu den erzielten Ergebnissen gesetzt werden. Diese Kennzahlen werden in entsprechenden Tabellen für alle Gütekriterien dargestellt.

In Tabelle 16 werden die Kennzahlen für den PFC zusammengefasst und durch den maximalen bzw. den mittleren Monte-Carlo-Standardfehler sowie dessen Median ergänzt. Die maximalen Monte-Carlo-Standardfehler sind für alle Verfahren in Tabelle 16 wesentlich

kleiner als die jeweiligen Ergebnisbereiche der Verfahren, was für die Verlässlichkeit der Ergebnisse spricht.

Verfahren	$\hat{\sigma}_{\text{MC,max}}$	$\hat{\sigma}_{\text{MC,mean}}$	$\hat{\sigma}_{\text{MC,median}}$	SP	$\hat{\sigma}_{\text{MC,max}} / SP$	$l_{\max}$
forest	0.002	0.00037	0.00028	0.21905	0.00915	0.00786
mode	0.00311	0.00023	0.00018	0.2268	0.01373	0.01221
HD	0.00108	0.00033	0.00028	0.20327	0.00533	0.00424
kNN	0.00211	0.00038	0.00028	0.21822	0.00969	0.00829
irmi	0.00138	0.00028	0.00022	0.22225	0.00621	0.00541
MCA	0.00197	0.00031	0.00025	0.22631	0.00869	0.00771
EM_det	0.00688	0.0005	0.00028	0.21129	0.03255	0.02696
EM_stoch	0.00653	0.00053	0.00031	0.20416	0.03198	0.02559
EM_cat	0.00224	0.00049	0.00038	0.20487	0.01091	0.00876

Tabelle 16: Übersicht: Monte-Carlo-Standardfehler für den PFC

Zudem sind die maximalen Konfidenzintervalle, welche um den Mittelwert mit dem höchsten Monte-Carlo-Standardfehler gezeichnet werden würden, deutlich kleiner als der entsprechende Ergebnisbereich. Einzelne Ausreißer der stochastischen und deterministischen EM-Imputationsverfahren werden anhand des Vergleichs vom Median mit dem Mittelwert erkennbar.

Für die Monte-Carlo-Standardfehler der Verteilungsabweichung ergibt sich ein sehr ähnliches Bild. Die entsprechenden Kennzahlen sind in Tabelle 17 zu sehen. In der Tabelle wird deutlich, dass das Verhältnis zwischen dem maximalen Monte-Carlo-Standardfehler und der Spannweite für alle Verteilungsabweichungen sehr gering ist, was wiederum für die Verlässlichkeit dieser Ergebnisse spricht.

Verfahren	$\hat{\sigma}_{\text{MC,max}}$	$\hat{\sigma}_{\text{MC,mean}}$	$\hat{\sigma}_{\text{MC,median}}$	SP	$\hat{\sigma}_{\text{MC,max}} / SP$	$l_{\max}$
forest	0.00251	0.00045	0.00033	0.16274	0.01541	0.00983
mode	0.00311	0.00023	0.00018	0.2268	0.01373	0.01221
HD	0.00166	0.00039	0.00032	0.13001	0.01274	0.00649
kNN	0.0023	0.00042	0.00033	0.13636	0.01687	0.00902
irmi	0.00226	0.0004	0.0003	0.17023	0.0133	0.00887
MCA	0.00218	0.00041	0.00032	0.1967	0.01106	0.00853
EM_det	0.01359	0.00068	0.00034	0.43368	0.03134	0.05328
EM_stoch	0.01021	0.00065	0.00035	0.40125	0.02545	0.04003
EM_cat	0.00474	0.00075	0.00059	0.20019	0.0237	0.0186

Tabelle 17: Übersicht: Monte-Carlo-Standardfehler für die Verteilungsabweichung

Ebenso verhält es sich mit den maximalen Längen der Konfidenzintervalle der einzelnen Verfahren, auch diese sind gemessen an der Spannweite des Gütekriteriums wesentlich geringer. Die vereinzelt erhöhten Werte spiegeln sich im Vergleich der Mediane mit den durchschnittlichen Monte-Carlo-Standardfehlern wider.

Auch die Monte-Carlo-Standardfehler der Korrelationsschätzung verhalten sich sehr ähnlich zu den zuvor betrachteten Monte-Carlo-Standardfehlern. Die Tabelle 18 gibt weitere Informationen bezüglich der Sicherheit dieser Ergebnisse. Sowohl die Quotienten aus den maximalen Monte-Carlo-Standardfehlern und den Spannweiten als auch die maximalen Längen der Konfidenzintervalle sind in diesem Fall sehr gering und sprechen für die Verlässlichkeit der Ergebnisse aller Verfahren. Hier ist nun auch bei fast allen Verfahren eine Abweichung des Medians vom Mittelwert, bedingt durch teilweise leicht erhöhte Monte-Carlo-Standardfehler, erkennbar.

Verfahren	$\hat{\sigma}_{\text{MC,max}}$	$\hat{\sigma}_{\text{MC,mean}}$	$\hat{\sigma}_{\text{MC,median}}$	SP	$\hat{\sigma}_{\text{MC,max}} / SP$	$l_{\max}$
forest	0.00251	0.00045	0.00033	0.16274	0.01541	0.00983
mode	0.00311	0.00023	0.00018	0.2268	0.01373	0.01221
HD	0.00166	0.00039	0.00032	0.13001	0.01274	0.00649
kNN	0.0023	0.00042	0.00033	0.13636	0.01687	0.00902
irmi	0.00226	0.0004	0.0003	0.17023	0.0133	0.00887
MCA	0.00218	0.00041	0.00032	0.1967	0.01106	0.00853
EM_det	0.01359	0.00068	0.00034	0.43368	0.03134	0.05328
EM_stoch	0.01021	0.00065	0.00035	0.40125	0.02545	0.04003
EM_cat	0.00474	0.00075	0.00059	0.20019	0.0237	0.0186

Tabelle 18: Übersicht: Monte-Carlo-Standardfehler für die Korrelationsschätzung

Die Monte-Carlo-Standardfehler für die RMSE der Regressionskoeffizienten sind sehr gering. Lediglich für den Fall einer ungleichen Verteilung mit wenigen Ausprägungen werden geringe Erhöhungen deutlich. Die Berechnung der Kennwerte würde durch diese Verteilung verzerrt werden. Zudem hat eine genauere Betrachtung gezeigt, dass auch hier die maximalen Längen der Konfidenzintervalle verhältnismäßig hoch sind. Aus diesem Grund erfolgt die nachfolgende Betrachtung ohne die zuvor beschriebene Verteilung mit erhöhten Monte-Carlo-Standardfehlern. Eine Übersicht dazu ist in Tabelle 19 zu sehen. Auch hier erscheinen die Konfidenzintervalle im Verhältnis zur Spannweite deutlich höher als für die

zuvor betrachteten Gütekriterien, jedoch handelt es sich bei den maximalen Monte-Carlo-Standardfehlern um vereinzelte Werte, die im Fall des MNAR-Ausfallmechanismus auftreten können.

Verfahren	$\hat{\sigma}_{\text{MC,max}}$	$\hat{\sigma}_{\text{MC,mean}}$	$\hat{\sigma}_{\text{MC,median}}$	SP	$\hat{\sigma}_{\text{MC,max}} / SP$	l_{max}
forest	0.0749	0.0085	0.0066	1.9545	0.0383	0.2936
mode	0.101	0.0082	0.0063	1.9764	0.0511	0.3961
HD	0.0425	0.0068	0.0061	1.4922	0.0285	0.1667
kNN	0.0522	0.0072	0.0063	1.6083	0.0324	0.2045
irmi	0.101	0.0081	0.006	1.8238	0.0554	0.3959
MCA	0.0751	0.0081	0.0065	2.0631	0.0364	0.2943
EM_det	0.0876	0.0082	0.0059	1.7756	0.0493	0.3435
EM_stoch	0.0858	0.0075	0.0061	1.4714	0.0583	0.3363
EM_cat	0.0676	0.0075	0.0065	0.5904	0.1145	0.2649

Tabelle 19: Übersicht: Monte-Carlo-Standardfehler für die Regressionskoeffizienten

Dies fällt insbesondere beim Vergleich der maximalen Standardfehler mit dem Durchschnitt bzw. dem Median derselben auf. Demzufolge können die Ergebnisse auch hier als gesichert betrachtet werden, denn nichtsdestotrotz sind auch die Quotienten aus den maximalen Monte-Carlo-Standardfehler und den verfahrensspezifischen Spannweiten sehr gering.

4 Fazit

Nachdem gezeigt werden konnte, dass nahezu alle Ergebnisse als gesichert betrachtet werden können, sollen diese nun zusammengefasst werden. Die Simulationsstudie hat gezeigt, dass kein Verfahren über alle Faktorstufenkombinationen hinweg bezüglich eines Gütekriteriums die besten Ergebnisse erzielt. Außerdem konnte festgestellt werden, dass die Ergebnisse zwischen den Gütekriterien und innerhalb dieser sehr stark variieren können. Deshalb werden nachfolgend die Empfehlungen für Imputationsverfahren in Abhängigkeit vom jeweiligen Gütekriterium und den entsprechenden Faktorstufen gegeben. In Anlehnung an Rockel (2022, S. 229 ff.) werden Entscheidungsbäume mit dem R-Paket rpart mittels CART-Algorithmus erzeugt (vgl. Therneau et al., 2022). Im Gegensatz zu Rockel (2022) wird aus Gründen der Übersichtlichkeit für jedes Gütekriterium ein eigenständiger

Entscheidungsbaum erstellt. Dies ermöglicht die Auswahl des besten Verfahrens für ein Gütekriterium. Dazu werden zuvor die besten Verfahren jeder Teilabbildung bestimmt, indem die Mittelwerte über alle Ausfallraten für jedes Verfahren gebildet und miteinander verglichen werden. Anschließend kann die Abhängigkeit des besten Verfahrens von den restlichen Faktorstufen bestimmt werden.

Der Entscheidungsbaum für das Gütekriterium PFC ist in Abbildung 10 zu sehen. Hier ist direkt erkennbar, dass die Anzahl der Objekte nur einen geringfügigen Einfluss auf das Gütekriterium besitzt und somit nicht Teil des Entscheidungsbaums ist. Für den Fall, dass der Ausfallmechanismus MCAR oder MAR ist, trifft die Wahl überwiegend auf MCA. Es gibt demnach Ausnahmen, in denen andere Verfahren bessere Ergebnisse erzeugen als MCA. Dies ist beispielsweise der Fall, wenn die Korrelation gering ist und die Verteilung unausgeglichen. Für diese Konstellation wird die Modusimputation stark bevorzugt, diese ist hier die Wahl des Verfahrens. Ebenso liefert irmi bessere Ergebnisse, wenn die Verteilung unausgeglichen mit wenigen Ausprägungen und die Korrelation 0,3 oder 0,7 ist.

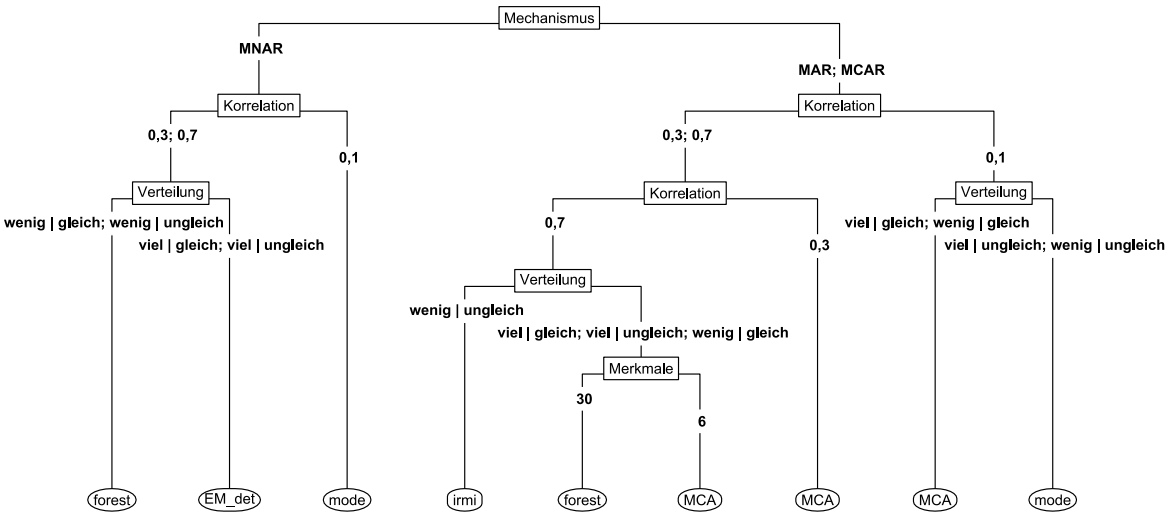

Abbildung 10: Entscheidungsbaum PFC

Für alle anderen Verteilungsformen ist bei gleichen Korrelationen die Anzahl der Merkmale entscheidend. Sind diese hoch, ist missForest als Verfahren zu empfehlen. Im Fall eines MNAR-Ausfallmechanismus ist bei einer geringen Korrelation immer die Modusimputati-

on zu wählen, bei höheren Korrelationen für wenige Ausprägungen missForest und für viele Ausprägungen die deterministische EM-Imputation.

In Abbildung 11 ist der Entscheidungsbaum für die Verteilungsabweichung dargestellt.

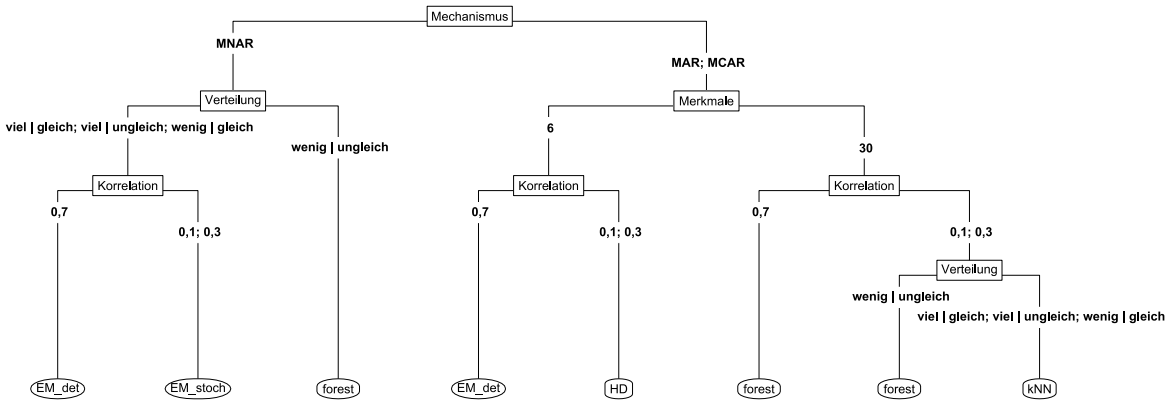

Abbildung 11: Entscheidungsbaum Verteilungsabweichung

Wie im Fall für PFC wird der geringe Einfluss durch die Anzahl der Objekte dadurch erkennbar, dass keine Entscheidung nach dieser getroffen wird. Das Verfahren, welches am häufigsten die besten Ergebnisse erzielt, ist in diesem Fall missForest. Demnach wird insbesondere auf die Entscheidungen eingegangen, wenn missForest nicht das beste Verfahren ist. Für die Ausfallmechanismen MCAR und MAR ist im Fall weniger Merkmale bei einer hohen Korrelation das deterministische EM-Imputationsverfahren am besten. Wenn jedoch die Korrelation nicht hoch ist, also 0,1 oder 0,3, dann liefert Random Hot-Deck die besten Ergebnisse. Unter denselben Ausfallmechanismen erweist sich kNN als das beste Verfahren in Situationen mit vielen Merkmalen, mit mittlerer bis geringer Korrelation und mit einer Verteilung, die weder unausgeglichen ist, noch wenige Ausprägungen aufweist. Unter dem MNAR-Ausfallmechanismus ist missForest nicht das beste Verfahren, wenn keine unausgeglichene Verteilung mit wenigen Ausprägungen vorliegt. Ist dies der Fall, entscheidet es sich zwischen der deterministischen und der stochastischen EM-Imputation. Erstere ist zu empfehlen, wenn $\rho = 0,7$ ist. Fällt diese geringer aus, so wird letztere empfohlen.

Der Entscheidungsbaum für das beste Verfahren zur Wiederherstellung der ursprünglichen Korrelation ist in Abbildung 12 zu sehen. Es fällt auf, dass hier die Anzahl der Merkmale nicht von Bedeutung ist, dafür aber erstmals die Objektanzahl. Eine erste Entscheidung wird nun zwischen den verschiedenen Stufen der Korrelation getroffen, was wie zuvor beschrieben an der damit verbunden Zielkorrelation liegt, welche es wiederherzustellen gilt.

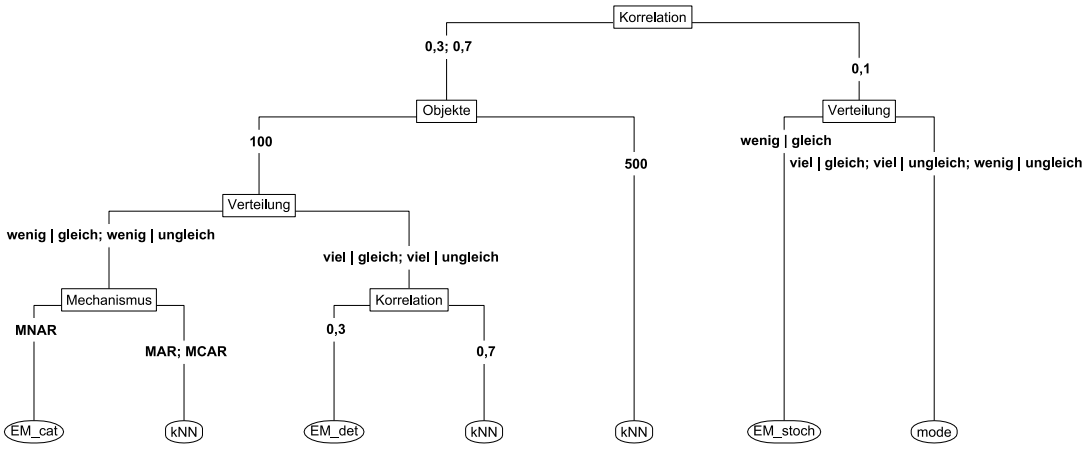

Abbildung 12: Entscheidungsbaum Korrelation

Das beste Verfahren ist am häufigsten ist die Modusimputation. Das wird auch im Entscheidungsbaum daran erkennbar, dass diese immer das beste Verfahren ist, wenn die Korrelation $\rho = 0,1$ ist, außer die Verteilung ist ausgeglichen mit wenigen Ausprägungen. In diesem Fall ist die stochastische EM-Imputation das beste Verfahren. Für höhere Korrelationen ist kNN immer das geeignetste Verfahren, wenn die Objektanzahl $n = 500$ ist. Ist diese nicht hoch, wird nach Verteilungen mit wenigen und vielen Ausprägungen unterschieden. Im Fall weniger Ausprägungen ist kNN für die Ausfallmechanismen MCAR und MAR zu empfehlen, für den MNAR-Ausfallmechanismus ist es die EM-Imputation speziell für kategoriale Daten. Liegen viele Ausprägungen vor, ist kNN das beste Verfahren, wenn $\rho = 0,7$ ist. Diese Position wird jedoch durch die deterministische EM-Imputation eingenommen, falls $\rho = 0,3$ ist.

Der Entscheidungsbaum für die logistische Regression ist kritisch zu betrachten. Dieser ist in Abbildung 13 dargestellt. Das ist der Sicherheit der Ergebnisse für die ungleiche Verteilung mit wenigen Ausprägungen geschuldet. Der entsprechende Pfad benötigt keine weitere Betrachtung, da die Ergebnisse hier zu unsicher sind. Er ist dennoch der Vollständigkeit halber dargestellt. Abgesehen davon liefert für den MAR- und den MCAR-Ausfallmechanismus sowie $\rho = 0{,}1$ die Modusimputation die besten Ergebnisse. Für $\rho \neq 0{,}1$ ist es in dem Fall MCA. Für den MNAR-Ausfallmechanismus liefert Random Hot Deck für $\rho = 0{,}1$ die besten Ergebnisse, sonst ist es die deterministische EM-Imputation. Die Anzahl der Objekte und Merkmale sowie die Verteilung[14] besitzen keine Bedeutung für die Wahl des besten Verfahrens.

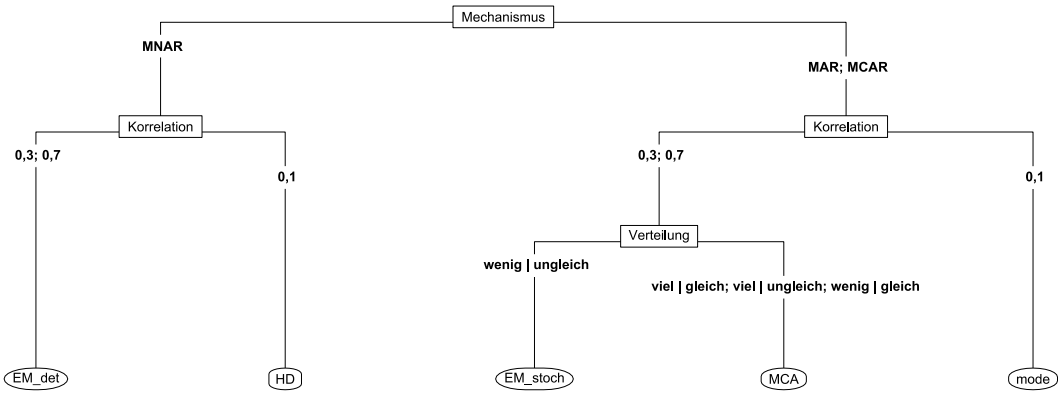

Abbildung 13: Entscheidungsbaum logistische Regression

Die hier gegebenen Entscheidungshilfen können nur auf Datenmatrizen mit vergleichbaren Faktorstufen angewendet werden. Es wäre daher von Interesse auch andere bzw. weitere Faktorstufen in Forschungsarbeiten zu berücksichtigen. Beispielsweise könnte die Betrachtung anderer Formen der Verteilung ebenso interessante Ergebnisse liefern wie eine noch höhere Anzahl an Objekten und Merkmalen. Es können auch Charakteristiken eines Datensatzes von Bedeutung sein, welche noch gar nicht in dieser Simulationsstudie betrachtet wurden. Darüber hinaus gilt es die Auswahl eines geeigneten Imputationsverfahrens stets kritisch zu hinterfragen und auf die entsprechende Eignung zu prüfen. Eine Evaluierung der Güte dieser Entscheidungshilfen stellt ebenfalls einen interessanten Punkt dar.

[14] Da die Ergebnisse für die ungleiche Verteilung mit wenigen Ausprägungen nicht als gesichert gelten.

5 Literaturverzeichnis

Agresti, Alan (2013): Categorical data analysis. 3. ed. Hoboken, NJ: Wiley-Interscience (Wiley series in probability and statistics).

Agresti, Alan (2019): An introduction to categorical data analysis. Third edition. Hoboken NJ: John Wiley & Sons (Wiley series in probability and statistics).

Andridge, Rebecca R.; Little, Roderick J. A. (2010): A Review of Hot Deck Imputation for Survey Non-response. In: *International Statistical Review* 78 (1), S. 40–64. DOI: 10.1111/j.1751-5823.2010.00103.x.

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

Cugnata, Federica; Salini, Silvia (2017): Comparison of alternative imputation methods for ordinal data. In: *Communications in Statistics - Simulation and Computation* 46 (1), S. 315–330. DOI: 10.1080/03610918.2014.963611.

Ferrari, Pier Alda; Annoni, Paola; Barbiero, Alessandro; Manzi, Giancarlo (2011): An imputation method for categorical variables with application to nonlinear principal component analysis. In: *Computational Statistics & Data Analysis* 55 (7), S. 2410–2420. DOI: 10.1016/j.csda.2011.02.007.

Fialkowski, Allison; Tiwari, Hemant (2019): SimCorrMix: Simulation of Correlated Data with Multiple Variable Types Including Continuous and Count Mixture Distributions. In: *The R Journal* 11 (1), S. 250. DOI: 10.32614/RJ-2019-022.

Fialkowski, Allison Cynthia (2022): SimCorrMix. Simulation of Correlated Data with Multiple Variable Types. Version 0.1.1. URL: https://CRAN.R-project.org/package=SimCorrMix.

Harding, Ted; Tusell, Fernando; Schafer, Joseph L. (2012): cat: Analysis of categorical-variable datasets with missing values. Analysis of categorical-variable datasets with missing values. Version 0.0-6.5. URL: https://CRAN.R-project.org/package=cat.

Husson, François; Josse, Julie (2020): missMDA. Handling Missing Values with Multivariate Data Analysis. Version 1.18. URL: https://CRAN.R-project.org/package=missMDA.

James, Gareth; Witten, Daniela; Hastie, Trevor; Tibshirani, Robert (2017): An introduction to statistical learning. With applications in R. Corrected at 8th printing. New York, Heidelberg, Dordrecht, London: Springer (Springer Texts in Statistics).

Josse, Julie; Husson, François (2016): missMDA : A Package for Handling Missing Values in Multivariate Data Analysis. In: *J. Stat. Soft.* 70 (1). DOI: 10.18637/jss.v070.i01.

Kowarik, Alexander; Templ, Matthias (2016): Imputation with the R Package VIM. In: *J. Stat. Soft.* 74 (7). DOI: 10.18637/jss.v074.i07.

Lin, Wei-Chao; Tsai, Chih-Fong (2020): Missing value imputation: a review and analysis of the literature (2006–2017). In: *Artif Intell Rev* 53 (2), S. 1487–1509. DOI: 10.1007/s10462-019-09709-4.

Little, Roderick J. A.; Rubin, Donald B. (2020): Statistical analysis with missing data. Third edition. Hoboken, NJ: John Wiley and Sons, Inc.; Wiley (Wiley series in probability and statistics).

Meyer, D.; Zeileis, A.; Hornik, K.; Gerber, F.; Friendly, M. (2022): vcd: Visualizing Categorical Data. Version 1.4-10. URL: https://CRAN.R-project.org/package=vcd.

Morris, Tim P.; White, Ian R.; Crowther, Michael J. (2019): Using simulation studies to evaluate statistical methods. In: *Statistics in medicine* 38 (11), S. 2074–2102. DOI: 10.1002/sim.8086.

R Core Team (2022): R: A Language and Environment for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Rhemtulla, Mijke; Brosseau-Liard, Patricia É.; Savalei, Victoria (2012): When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. In: *Psychological Methods* 17 (3), S. 354–373. DOI: 10.1037/a0029315.

Rockel, Tobias (2020): missMethods. Methods for Missing Data. Version 0.2.0. URL: https://CRAN.R-project.org/package=missMethods.

Rockel, Tobias (2022): Güteuntersuchung von Imputationsverfahren für unvollständige Datenmatrizen. Universitätsverlag Ilmenau.

Röhrig, Steve; Rockel, Tobias (2020): Analyse existierender Simulationsstudien zum Umgang mit fehlenden qualitativen Daten. Ilmenau (Ilmenauer Beiträge zur Wirtschaftsinformatik). URL: https://nbn-resolving.org/urn:nbn:de:gbv:ilm1-2020200439.

Schafer, Joseph L. (1997): Analysis of Incomplete Multivariate Data. 1. ed. Boca Raton, Fla.: Chapman & Hall/CRC (Monographs on statistics and applied probability, 72).

Stekhoven, Daniel J. (2022): missForest. Nonparametric Missing Value Imputation using Random Forest. Version 1.5. URL: https://CRAN.R-project.org/package=missForest.

Stekhoven, Daniel J.; Bühlmann, Peter (2012): MissForest--non-parametric missing value imputation for mixed-type data. In: *Bioinformatics (Oxford, England)* 28 (1), S. 112–118. DOI: 10.1093/bioinformatics/btr597.

Templ, Matthias; Kowarik, Alexander; Alfons, Andreas; Cillia, Gregor de; Rannetbauer, Wolfgang (2021): VIM. Visualization and Imputation of Missing Values. Version 6.1.1. URL: https://CRAN.R-project.org/package=VIM.

Therneau, Terry; Atkinson, Beth; Ripley, Brian (2022): rpart. Recursive Partitioning and Regression Trees. Version 4.1.19. URL: https://CRAN.R-project.org/package=rpart.

Troyanskaya, O.; Cantor, M.; Sherlock, G.; Brown, P.; Hastie, T.; Tibshirani, R. et al. (2001): Missing value estimation methods for DNA microarrays. In: *Bioinformatics (Oxford, England)* 17 (6), S. 520–525. DOI: 10.1093/bioinformatics/17.6.520.

Vidotto, Davide; Vermunt, Jeroen K.; van Deun, Katrijn (2018): Bayesian Latent Class Models for the Multiple Imputation of Categorical Data. In: *Methodology* 14 (2), S. 56–68. DOI: 10.1027/1614-2241/a000146.

Wu, Wei; Jia, Fan; Enders, Craig (2015): A Comparison of Imputation Strategies for Ordinal Missing Data on Likert Scale Variables. In: *Multivariate behavioral research* 50 (5), S. 484–503. DOI: 10.1080/00273171.2015.1022644.