Ilmenauer Beiträge zur Wirtschaftsinformatik

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

Tobias Rockel

Vergleich von Imputationsverfahren – Eine Simulationsstudie

Arbeitsbericht Nr. 2018-01, November 2018

Technische Universität Ilmenau
Fakultät für Wirtschaftswissenschaften und Medien
Institut für Wirtschaftsinformatik

Autor: Tobias Rockel

Titel: Vergleich von Imputationsverfahren – Eine Simulationsstudie

Ilmenauer Beiträge zur Wirtschaftsinformatik Nr. 2018-01, Technische Universität Ilmenau, 2018

ISSN 1861-9223

ISBN 978-3-938940-60-0

URN urn:nbn:de:gbv:ilm1-2018200160

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

Gliederung

Zusammenfassung: Imputationsverfahren stellen eine Strategie zum Umgang mit fehlenden Werten dar. Der Vorteil der Imputationsverfahren ist, dass sie einen vervollständigten Datensatz zur Verfügung stellen, der mit Hilfe herkömmlicher Analyseverfahren ausgewertet werden kann. Dabei ist jedoch zu beachten, dass eine Imputation auch zu Verzerrungen von Analyseergebnissen führen kann. Die Auswirkungen verschiedener Imputationsverfahren auf unterschiedliche Datensatzstrukturen wird in diesem Arbeitspapier mit Hilfe einer Simulationsstudie untersucht. Dabei zeigt sich, dass keins der untersuchten Imputationsverfahren in jeder Situation allen anderen Verfahren überlegen ist. Vielmehr sollte bei der Auswahl eines Imputationsverfahrens das Analyseziel und die Struktur des vorliegenden Datensatzes berücksichtigt werden.

Schlüsselworte: Imputation, fehlende Werte, Simulationsstudie

1 Einführung

Die Anwendung der meisten Datenanalyseverfahren setzt voraus, dass eine Datenmatrix ohne fehlende Werte vorliegt (vgl. Schafer und Graham, 2002, S. 147). Diese Voraussetzung ist bei empirischen Untersuchungen häufig jedoch nicht erfüllt. So stellen zum Beispiel Eekhout et al. (2012, S. 729–731) bei einer Untersuchung von 285 epidemiologischen Studien fest, dass maximal 8 % der Studien auf vollständige Datensätze zurückgreifen konnten. Backhaus und Blechschmidt (2009, S. 266) gehen sogar so weit, dass sie behaupten, es gäbe in der Realität praktisch keinen Datensatz ohne fehlende Werte.

Aus diesem Grund ist es im Vorfeld der eigentlichen Datenanalyse meist notwendig, zunächst eine Strategie zum Umgang mit fehlenden Werten zu wählen. Eine der einfachsten Strategien ist, die fehlenden Werte durch Schätzwerte zu ersetzen. Verfahren, die die fehlende Werte durch Schätzwerte ersetzen, werden Imputationsverfahren genannt. Der Vorteil dieser Verfahren ist, dass aus ihrer Anwendung ein vervollständigter Datensatz resultiert, der mit herkömmlichen Analyseverfahren ausgewertet werden kann. Jedoch kann die unreflektierte Verwendung von Imputationsverfahren die Analyseergebnisse verzerren. Diese Verzerrung ist unter anderem von dem gewählten Imputationsverfahren und dem Ausfallmechanismus im Datensatz abhängig (vgl. Bankhofer, 1995, S. 104–105).

Der Ausfallmechanismus beschreibt den Zusammenhang zwischen dem Fehlen der Werte und den Werten in der Datenmatrix. Konkreter geht es um die stochastische Abhängigkeit zwischen der Missing Data (MD) Indikatormatrix, welche das Fehlen bzw. Vorhandensein der Werte in der Datenmatrix anzeigt, und der Datenmatrix A. Für die Definition der Ausfallmechanismen werden sowohl die MD-Indikatormatrix als auch die Datenmatrix A als Zufallsvariablen aufgefasst. Falls die MD-Indikatormatrix und die Datenmatrix A stochastisch unabhängig sind, werden die Daten als Missing Completely at Random (MCAR) bezeichnet. Der MCAR Ausfallmechanismus ist ein Speziallfall des Missing at Random (MAR) Ausfallmechanismus, bei dem die MD-Indikatormatrix von den beobachteten Werten, aber nicht von den unbeobachteten Werten abhängen darf. Falls das Fehlen der Werte auch von unbeobachteten Werten abhängt, sind die Daten Not Missing at Random (NMAR) (vgl. Little und Rubin, 2002, S. 12).

Im Rahmen dieses Arbeitspapiers wird die Eignung verschiedener Imputationsverfahren bei unterschiedlichen Datenmatrizen und Ausfallszenarien untersucht. Dazu wird

eine Simulationsstudie durchgeführt, deren Aufbau im Abschnitt 2 beschrieben ist. Anschließend werden im Abschnitt 3 die Ergebnisse der Simulationsstudie dargestellt. Diese werden im Abschnitt 4 zusammengefasst und interpretiert. Abschließend werden im Abschnitt 5 noch ein Fazit und Handlungsempfehlungen aus den Simulationsergebnissen abgeleitet.

2 Design der Simulationsstudie

In diesem Abschnitt wird das Design der durchgeführten Studie dargestellt. Zunächst wird auf die Erzeugung der Datenmatrizen eingegangen. Anschließend werden die verwendeten Ausfallmechanismen und Imputationsverfahren vorgestellt. Zum Abschluss werden die Bewertungskriterien für die MD-Verfahren erläutert. Der Studie liegt ein vollständiger Versuchsplan zugrunde, das heißt, es werden alle mögliche Faktorkombination der untersuchten Einflussfaktoren simuliert. Die Durchführung der Simulationsstudie erfolgt mittels der Statistikprogrammiersprache R (R Core Team, 2018) in der Version 3.5.0.

Aus einer vorherigen Betrachtung (Rockel, 2017, S. 25–26) geht hervor, dass die Anzahl der Objekte und der Merkmale zwei wichtige Einflussfaktoren bei der Auswahl eines MD-Verfahren sind. Daher werden in der Simulationsstudie Datensätze mit $n = 100$ und $n = 500$ Objekten erzeugt. Ferner wird die Anzahl der Merkmale zwischen den beiden Stufen $m = 5$ und $m = 25$ variiert. Neben der reinen Objekt- und Merkmalsanzahl hat auch der Zusammenhang zwischen den Merkmalen einen Einfluss auf viele MD-Verfahren. Daher werden Datensätze mit drei unterschiedlichen Korrelationsstärken simuliert. Die Korrelationsstufen $\rho = 0{,}1; 0{,}4; 0{,}7$ repräsentieren die Fälle, dass ein sehr schwacher, ein mittlerer oder ein starker Zusammenhang zwischen den Merkmalen im Datensatz vorliegt. Bei den simulierten Datensätze ist die theoretische Korrelation zwischen allen Merkmalen gleich hoch.

Zur Erzeugung der Datenmatrizen wird eine multivariate Normalverteilung mit einem Erwartungswertvektor $\mu = 0$ verwendet. Ferner wird die Varianz in allen Merkmalen auf Eins gesetzt. Die Kovarianz zwischen zwei Merkmalen entspricht dann der Korrelation ρ. Die Simulation der Datensätze in R geschieht mit Hilfe des Pakets mvtnorm (Genz, Bretz et al., 2018) in der Version 1.0.7, welches auf dem Buch von Genz und Bretz (2009) basiert.

Zur Erzeugung der fehlenden Werte werden sowohl ein MCAR als auch zwei MAR Ausfallmechanismen verwendet. Dabei wird ein multivariates Ausfallmuster in den letzten $\lfloor \frac{m}{2} \rfloor$ Merkmalen $a_{\lceil \frac{m}{2}+1 \rceil}, a_{\lceil \frac{m}{2}+2 \rceil}, \ldots, a_m$ eines Datensatzes erzeugt. Hierbei rundet die Funktion $\lfloor x \rfloor$ bzw. $\lceil x \rceil$ die Zahl x ab bzw. auf. Ferner wird der Anteil der fehlenden Werte in den Variablen, die vom Ausfall betroffen sind, von 10 % bis 50 % in 10 %-Schritten variiert. Für den MCAR Ausfallmechanismus werden in jeder Variable mit fehlenden Werten zufällig Werte gelöscht. Die Anzahl der gelöschten Werte in einem Merkmal mit fehlenden Werten ist dabei stets $p \cdot n$, $p = 0{,}1; 0{,}2; \ldots; 0{,}5$.

Für den ersten MAR Ausfallmechanismus wird zunächst für jedes Merkmal mit fehlenden Werten ein anderes Merkmal ausgewählt, das den Ausfall steuert. Die Zuordnung ist schematisch in der Abbildung 1 dargestellt. Die Abbildung zeigt, dass das Merkmal $a_{\lceil \frac{m}{2} \rceil}$ den Ausfall im Merkmal $a_{\lceil \frac{m}{2}+1 \rceil}$, das Merkmal $a_{\lceil \frac{m}{2}-1 \rceil}$ den Ausfall im Merkmal $a_{\lceil \frac{m}{2}+2 \rceil}$ usw. steuert. Die Darstellung in der Abbildung 1 geht davon aus, dass die Merkmalsanzahl m ungerade ist. In diesem Fall steuert das zweite Merkmal den Ausfall im letzten Merkmal. Das erste Merkmal hingegen steuert keinen Ausfall, da nur in $\lfloor \frac{m}{2} \rfloor$ Merkmalen fehlende Werte erzeugt werden.

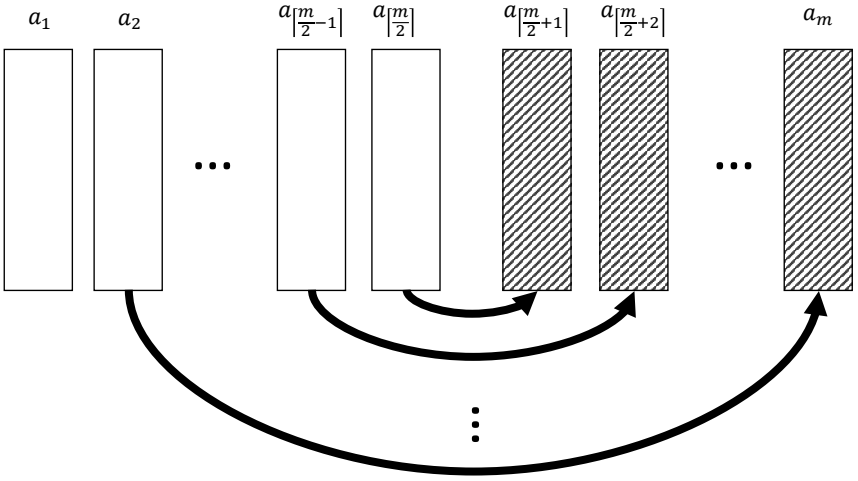

Abbildung 1: Definition des MAR Ausfallmechanismus (m ungerade)

Der MAR Ausfallmechanismus steuert den Ausfall in einem Merkmal k' anhand des Medians a_k^{med} im „ausfallsteuerenden" Merkmal k. Sei $I_{k,<med} = \{i \in \{1, \ldots, n\} \mid a_{ik} < a_k^{med}\}$ die Indexmenge der Objekte, deren Wert im Merkmal k kleiner ist als der Median, und $I_{k,\geq med} = \{i \in \{1, \ldots, n\} \mid a_{ik} \geq a_k^{med}\}$ die Indexmenge der Objekte, deren Wert

im Merkmal k größer-gleich dem Median ist. Für den Anteil $p = 0{,}1; 0{,}2; \ldots; 0{,}5$ an fehlenden Werte wählt der MAR Ausfallmechanismus $\frac{1}{3} \cdot p \cdot n$ Indizes aus $I_{k,<med}$ und $\frac{2}{3} \cdot p \cdot n$ Indizes aus $I_{k,\geq med}$ aus und löscht für diese ausgewählten Indizes die Werte im Merkmal k'. Das Verhältnis zwischen der Anzahl der fehlenden Werte in den Objekten, deren Ausprägungen im Merkmal k kleiner als der Median sind, zu der Anzahl der fehlenden Werte in den Objekten, deren Ausprägungen im Merkmal k größer-gleich dem Median sind, entspricht folglich 1:2. Deshalb wird der Ausfallmechanismus im Folgenden als MAR1:2 bezeichnet.

Der zweite MAR Ausfallmechanismus entspricht in der Struktur dem ersten. Er unterscheidet sich nur durch das gewählte Verhältnis von 1:4 anstatt von 1:2 vom ersten Ausfallmechanismus. Er wird daher als MAR1:4 bezeichnet. Der MAR1:4 Ausfallmechanismus stellt also eine Verstärkung des MAR1:2 Ausfallmechanismus dar, wodurch der Effekt verschieden starker MAR Ausfallmechanismen auf die MD-Verfahren untersucht werden kann. Ferner kann der verwendete MCAR Ausfallmechanismus als eine Art MAR1:1 Ausfallmechanismus interpretiert werden.[1] Deshalb können der MAR1:2 und der MAR1:4 Mechanismus auch als eine Verstärkung des MCAR Ausfallmechanismus interpretiert werden. Im Folgenden wird daher der MAR1:4 Ausfallmechanismus auch als stärkster und der MCAR Ausfallmechanismus als schwächster Ausfallmechanismus bezeichnet.

Zur Ersetzung der fehlenden Werte werden folgende acht Imputationsverfahren in der Studie untersucht:

- Mittelwertimputation (MW)
- deterministische Regressionsimputation (LR)
- stochastische Regressionsimputation (RR)
- adaptive Regressionsimputation (AR)
- Random Hot-Deck (RHD)
- Nearest-Neighbour Hot-Deck (NNHD)
- deterministische EM-Imputation (EMId)

[1] Es wird hier nur von einer Art MAR1:1 Mechanismus gesprochen, da der verwendete MCAR Ausfallmechanismus das Verhältnis nur im Mittel erfüllt, aber bei einzelnen Simulationsläufen das Verhältnis von 1:1 abweichen kann.

- stochastische EM-Imputation (EMIs)

Die Funktionen für die Mittelwertimputation, das Random Hot-Deck, die adpative Regressionsimputation (Bø et al., 2004) sowie die beiden EM-Imputationen werden in R implementiert. Die deterministische und die stochastische Regressionsimputation wird mittels des R-Pakets mice (van Buuren und Groothuis-Oudshoorn, 2011) in der Version 2.46.00 durchgeführt. Für das Nearest-Neighbour Hot-Deck wird die Funktion kNN aus dem R-Paket VIM (Kowarik und Templ, 2016) in der Version 4.7.0 verwendet.

Um die MD-Methoden zu bewerten, werden folgende Bewertungskriterien eingesetzt: Genauigkeit der Imputation, Schätzgüte des Erwartungswertvektors, der Varianzen und der Korrelationen. Die Abweichungen zwischen den wahren Werten bzw. Parametern und den imputierten Werten bzw. nach der Imputation geschätzten Parametern wird bei allen Kriterien mit Hilfe des Root Mean Squared Error (RMSE) berechnet. Für die Abweichungen zwischen der vervollständigten Datenmatrix A^{verv} und der Datenmatrix mit den Originalwerten A^{orig} berechnet sich der RMSE mittels

$$\sqrt{\frac{1}{n \cdot m} \sum_{i=1}^{n} \sum_{k=1}^{m} \left(a_{ik}^{verv} - a_{ik}^{orig} \right)^2}. \tag{1}$$

Anhand dieser Werte wird die Genauigkeit der Imputation bewertet.

Zur Bewertung der Schätzgüte des Erwartungswertes wird der RMSE zwischen dem wahren Erwartungswertvektor $\mu^{orig} = \left(\mu_1^{orig}, \ldots, \mu_m^{orig} \right)^T$ und dem geschätzten Erwartungswertvektor $\hat{\mu}^{verv} = (\hat{\mu}_1^{verv}, \ldots, \hat{\mu}_m^{verv})^T$ anhand des vervollständigten Datensatz berechnet:

$$\sqrt{\frac{1}{m} \sum_{k=1}^{m} \left(\hat{\mu}_k^{verv} - \mu_k^{orig} \right)^2} \tag{2}$$

Bei dem gewählten Simulationsdesign entspricht μ^{orig} entweder dem 5-dimensionalen Nullvektor (bei 5 Merkmalen) oder dem 25-dimensionalen Nullvektor (bei 25 Merkmalen), da alle Datensätze aus einer zentrierten multivariaten Normalverteilung stammen. Der Erwartungswert $\hat{\mu}_k^{verv}$ im Merkmal k wird anhand des Mittelwerts der vervollständigten Datenmatrix A^{verv} im Merkmal k geschätzt.

Analog wird für die Beurteilung der Varianzschätzung der RMSE zwischen den wahren Varianzen $\sigma^{2\,orig} = \left(\sigma_1^{2\,orig}, \ldots, \sigma_m^{2\,orig} \right)$ und den anhand des vervollständigten

Datensatzes geschätzten Varianzen $s^{2^{verv}} = \left(s_1^{2^{verv}}, \ldots, s_m^{2\ verv}\right)$ berechnet:

$$\sqrt{\frac{1}{m}\sum_{k=1}^{m}\left(s_k^{2^{verv}} - \sigma_k^{2^{orig}}\right)^2} \tag{3}$$

Da die Varianz in der Simulation für alle Merkmale mit Eins vorgegeben ist, ist $\sigma_k^{2^{orig}} = 1$ für alle $k \in \{1, \ldots, m\}$. Die empirische Varianz $s_k^{2^{verv}}$ wird anhand von allen Werten der vervollständigten Datenmatrix A^{verv} im Merkmal k geschätzt.

Die Abweichung zwischen der wahren Korrelation und der geschätzten Korrelation wird mittels

$$\sqrt{\frac{2}{m \cdot (m-1)}\sum_{k=1}^{m}\sum_{k<l}\left(r_{kl}^{verv} - \rho_{kl}^{orig}\right)^2} \tag{4}$$

berechnet. Dabei entspricht die wahre Korrelation ρ_{kl}^{orig} zwischen den Merkmalen k und l der aktuellen Faktorstufe $\rho = 0{,}1$, 0,4 oder 0,7 in der Simulation. Die geschätzte Korrelation r_{kl}^{verv} zwischen beiden Merkmalen wird mit Hilfe der Standardformel zur Korrelationsschätzung berechnet:

$$r_{kl}^{verv} = \frac{\sum_{i=1}^{n}\left(a_{ik}^{verv} - \bar{a}_{.k}^{verv}\right)\left(a_{il}^{verv} - \bar{a}_{.l}^{verv}\right)}{\sqrt{\sum_{i=1}^{n}\left(a_{ik}^{verv} - \bar{a}_{.k}^{verv}\right)^2 \sum_{i=1}^{n}\left(a_{il}^{verv} - \bar{a}_{.l}^{verv}\right)^2}}. \tag{5}$$

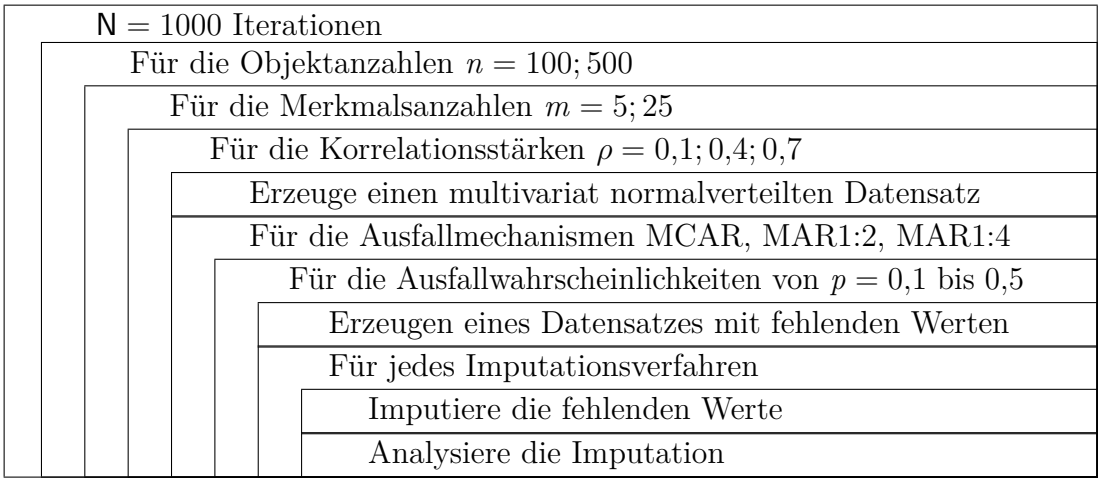

Abbildung 2: Das Design der Simulationsstudie

Für die Simulation wird jede Faktorstufenkombination 1000 Mal simuliert. Die im Folgenden dargestellten Ergebnisse sind jeweils Mittelwerte über diese N = 1000 Wieder-

holungen. Eine Übersicht über das Design der Simulationsstudie ist in der Abbildung 2 zu finden. Aus ihr geht hervor, dass insgesamt $2 \cdot 2 \cdot 3 \cdot 3 \cdot 5 = 180$ verschiedene Kombinationen aus Datensätzen und Ausfallszenarien in der Simulationsstudie betrachtet werden.

3 Ergebnisse der Simulationsstudie

3.1 Genauigkeit der Imputationswerte

Die Simulationsresultate zur Beurteilung der Imputationsgenauigkeit sind in den Abbildungen 3 und 4 dargestellt. Die Abbildung 3 enthält die Ergebnisse für $n = 100$ Objekte und die Abbildung 4 die Ergebnisse für $n = 500$ Objekte. Beide Abbildungen sind zeilenweise unterteilt in die drei im Abschnitt 2 beschriebenen Ausfallmechanismen MCAR, MAR1:2 und MAR1:4. In den ersten drei Spalten sind die Ergebnisse für $m = 5$ Merkmale und in den letzten drei Spalten die Ergebnisse für $m = 25$ Merkmale dargestellt. Ferner enthalten die erste und die vierte Spalte die Ergebnisse der Datensätze mit einer Korrelation von $\rho = 0{,}1$, die zweite und fünfte Spalte die Ergebnisse der Datensätze mit einer Korrelation von $\rho = 0{,}4$ und die dritte und sechste Spalte für eine Korrelation von $\rho = 0{,}7$.

Für jede Kombination wird auf der Abszissenachse der Anteil fehlender Werte sowie auf der Ordinatenachse der RMSE zwischen den imputierten Werten und den Originalwerten dargestellt. Je höher der RMSE ist, desto größer ist die Abweichung zwischen den Originalwerten und den imputierten Werten. Eine genaues Imputationsverfahren erzielt also kleinere RMSE-Werte als ein ungenaues Verfahren. Das genauste Verfahren ist folglich das Verfahren mit dem geringsten RMSE bei einer gegeben Faktorstufenkombination. Unter einer Faktorstufenkombination wird in diesem Zusammenhang die Kombination einer Datensatzstruktur (bestehend aus einer festgelegten Anzahl Objekten und Merkmalen sowie einer festgelegten Korrelation) mit einem Ausfallmechanismus und einem fixen Anteil fehlender Werte verstanden.

Bei den kleinsten Datensätzen ($n = 100$, $m = 5$), dargestellt auf der linken Seite der Abbildung 3, führen bei einer mittleren und hohen Korrelation meist die deterministische Regression- und Expectation Maximization (EM)-Imputation, dicht gefolgt von der adaptiven Regressionsimputation, zu den besten Ergebnissen. Wenn die Korrelation niedrig ist, liefert die Mittelwertimputation die genausten Imputationswerte, kann sich

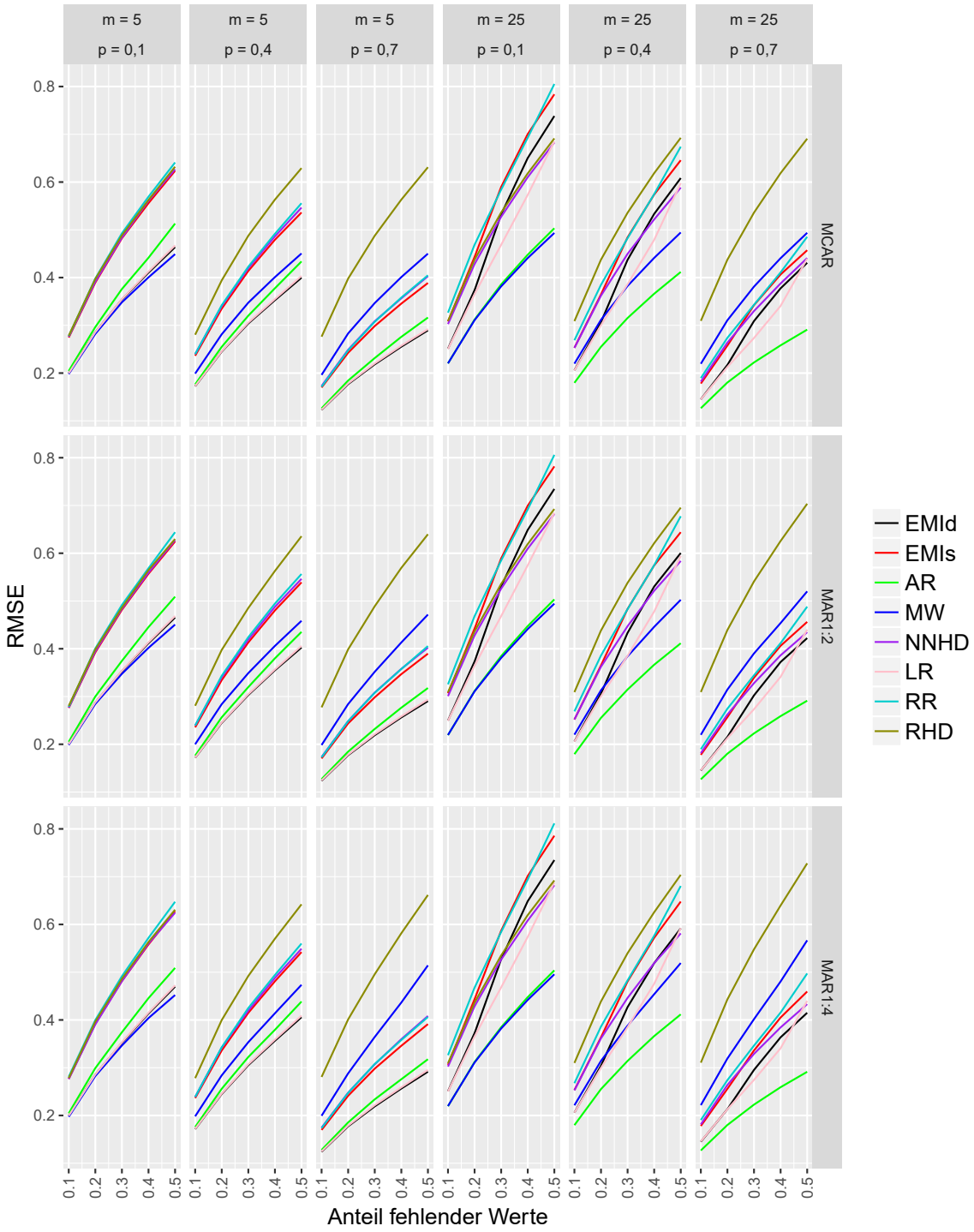

Abbildung 3: RMSE bei $n = 100$ Objekten zwischen Originalwerten und imputierten Werten

jedoch von den ersten drei Verfahren kaum abheben. Die Mittelwertimputation wird jedoch relativ zu den anderen drei Verfahren mit steigender Korrelation schlechter. Das Nearest-Neighbour Hot-Deck, die stochastische Regressions- und EM-Imputation liefern fast identische Genauigkeitswerte, die aber stets schlechter als die der ersten drei Verfahren sind. Das Random Hot-Deck ist für $\rho = 0{,}1$ noch mit diesen drei Verfahren vergleichbar, aber bei mittlerer und hoher Korrelation das mit Abstand schlechteste Verfahren.

Für die Datensätze mit wenigen Objekten und vielen Merkmalen ($n = 100$, $m = 25$), dargestellt im rechten Teil der Abbildung 3, stechen die Resultate der adaptiven Regressionsimputation hervor. Sie sind bei schwacher Korrelation $\rho = 0{,}1$ zusammen mit der Mittelwertimputation die besten und bei den anderen beiden Korrelationsstufen häufig mit Abstand die besten. Die stochastische Regressions- und EM-Imputation sind bei schwacher Korrelation die ungenausten Verfahren. Sie sind bei $\rho = 0{,}4$ aber besser als das Random Hot-Deck und können für $\rho = 0{,}7$ zusätzlich die Mittelwertimputation übertreffen. Die Ergebnisse des Nearest-Neighbour Hot-Decks entsprechen für einen niedrigen Anteil fehlender Werte in etwa den beiden stochastischen Verfahren, sind aber für höhere Anteile fehlender Werte besser als die der beiden stochastischen Verfahren. Das Random Hot-Deck ist für $\rho = 0{,}4$ und $\rho = 0{,}7$ das ungenauste Verfahren.

Bei den Datensätzen mit vielen Objekten ($n = 500$) und wenigen Merkmalen ($m = 5$, linke Seite der Abbildung 4) bilden die adaptive Regressionsimputation, die deterministische Regressionsimputation und die deterministische EM-Imputation die Gruppe mit den besten Ergebnissen über alle Korrelationsstufen. Innerhalb der Gruppe ist eine Differenzierung kaum möglich. Nur bei einem hohen Anteil fehlender Werte sind die Ergebnisse der adaptiven Regressionsimputation etwas schlechter als die der anderen beiden Verfahren. Eine weitere Dreiergruppe bildet das Nearest-Neighbour Hot-Deck, die stochastische Regressions- und EM-Imputation. Diese Dreiergruppe liefert bei der niedrigen Korrelation $\rho = 0{,}1$ zusammen mit dem Random Hot-Deck die ungenauesten Imputationswerte, kann sich jedoch für $\rho = 0{,}4$ von der Random Hot-Deck-Imputation absetzen und ist bei der höchsten Korrelationsstufe auch noch genauer als die Mittelwertimputation. Das Random Hot-Deck führt bei allen Datensätzen zu den ungenausten Imputationswerten. Die Mittelwertimputation liefert bei allen Korrelationsstufen ähnliche Genauigkeitswerte, verschlechtert sich dadurch aber bei der mittleren und hohen Korrelationen relativ zu den anderen Verfahren mit Ausnahme des Random Hot-Decks.

Die adaptive Regressionsimputation imputiert bei den größten Datensätzen ($n = 500$,

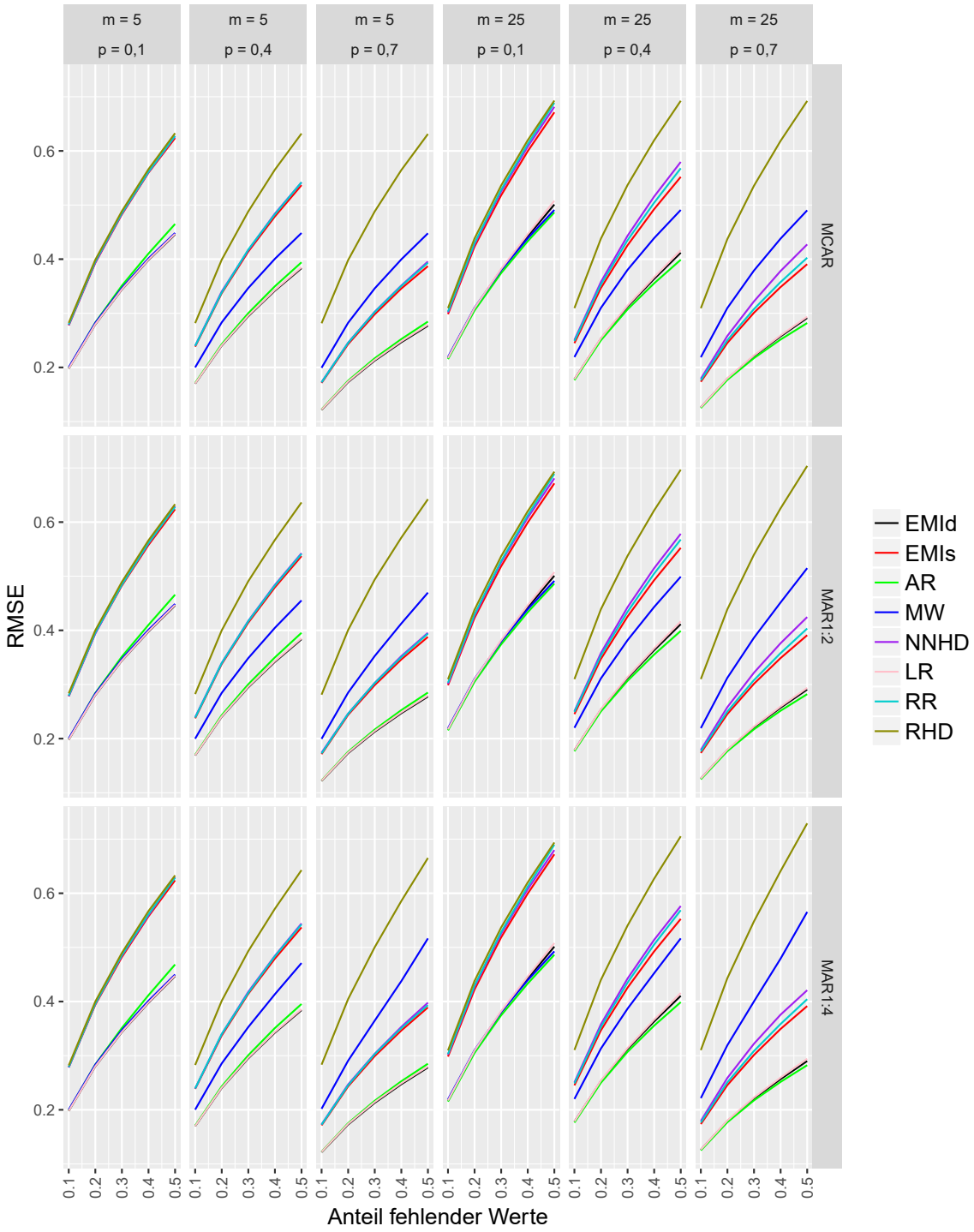

Abbildung 4: RMSE bei $n = 500$ Objekten zwischen Originalwerten und imputierten Werten

$m = 25$, rechter Teil der Abbildung 4) die Werte am genausten, oft dicht gefolgt von der deterministischen EM- und Regressionsimputation. Wie bei den Datensätzen mit vielen Objekten und wenigen Merkmalen bilden das Nearest-Neighbour Hot-Deck, die stochastische EM- und Regressionsimputation wieder eine Dreiergruppe, die sich mit zunehmender Korrelation verbessert und sich zunächst für $\rho = 0{,}4$ von dem Random Hot-Deck und für $\rho = 0{,}7$ zusätzlich von der Mittelwertimputation absetzen kann. Das Random Hot-Deck ist erneut über alle Korrelationsstufen das ungenauste Verfahren. Auch die Mittelwertimputation verhält sich analog zum Datensatz mit wenig Merkmalen und vielen Objekten.

Die Abbildungen 3 und 4 zeigen, dass alle Verfahren mit zunehmenden Anteil fehlender Werte ungenauer werden. Außerdem werden mit Ausnahme des Random Hot-Decks und der Mittelwertimputation alle Verfahren mit zunehmender Korrelation und Objektanzahl tendenziell genauer. Hingegen ist der Einfluss der Merkmalsanzahl auf die Verfahren nicht eindeutig. Die Steigerung des Ausfallmechanismus von MCAR über MAR1:2 hin zu MAR1:4 führt bei allen Verfahren zu einer leichten Verschlechterung der Ergebnisse. Die Genauigkeitseinbußen sind bei höherem Anteil fehlender Werte stärker als bei weniger fehlenden Werten. Ferner sind die Ergebnisse der Mittelwertimputation und des Random Hot-Deck stärker vom Ausfallmechanismus abhängig als die Ergebnisse der anderen Verfahren.

Insgesamt zeigt sich, dass die adaptive Regressionsimputation sowie die deterministische EM- und Regressionsimputation zu den genausten Imputationswerten führen. Des Weiteren liefert bei einer geringen Korrelation die Mittelwertimputation vergleichbare Ergebnisse zu diesen drei Verfahren, ist aber bei höherer Korrelation relativ gesehen schlechter. Die adaptive Regressionsimputation ist bei 25 Merkmalen tendenziell der deterministischen EM- und Regressionsimputation überlegen, während diese beiden Verfahren bei wenigen Merkmalen zu leicht besseren Ergebnissen als die adaptive Regressionsimputation führen.

3.2 Auswirkungen auf die Erwartungswertschätzung

Die Auswirkungen der Imputationsverfahren auf die Erwartungswertschätzung sind in den Abbildungen 5 und 6 dargestellt. Die beiden Abbildung sind analog zu den Abbildungen 3 und 4 aufgebaut. Der einzige Unterschied ist, dass nun auf der Ordinatenachse die Abweichung zwischen den wahren und den geschätzten Erwartungswerten

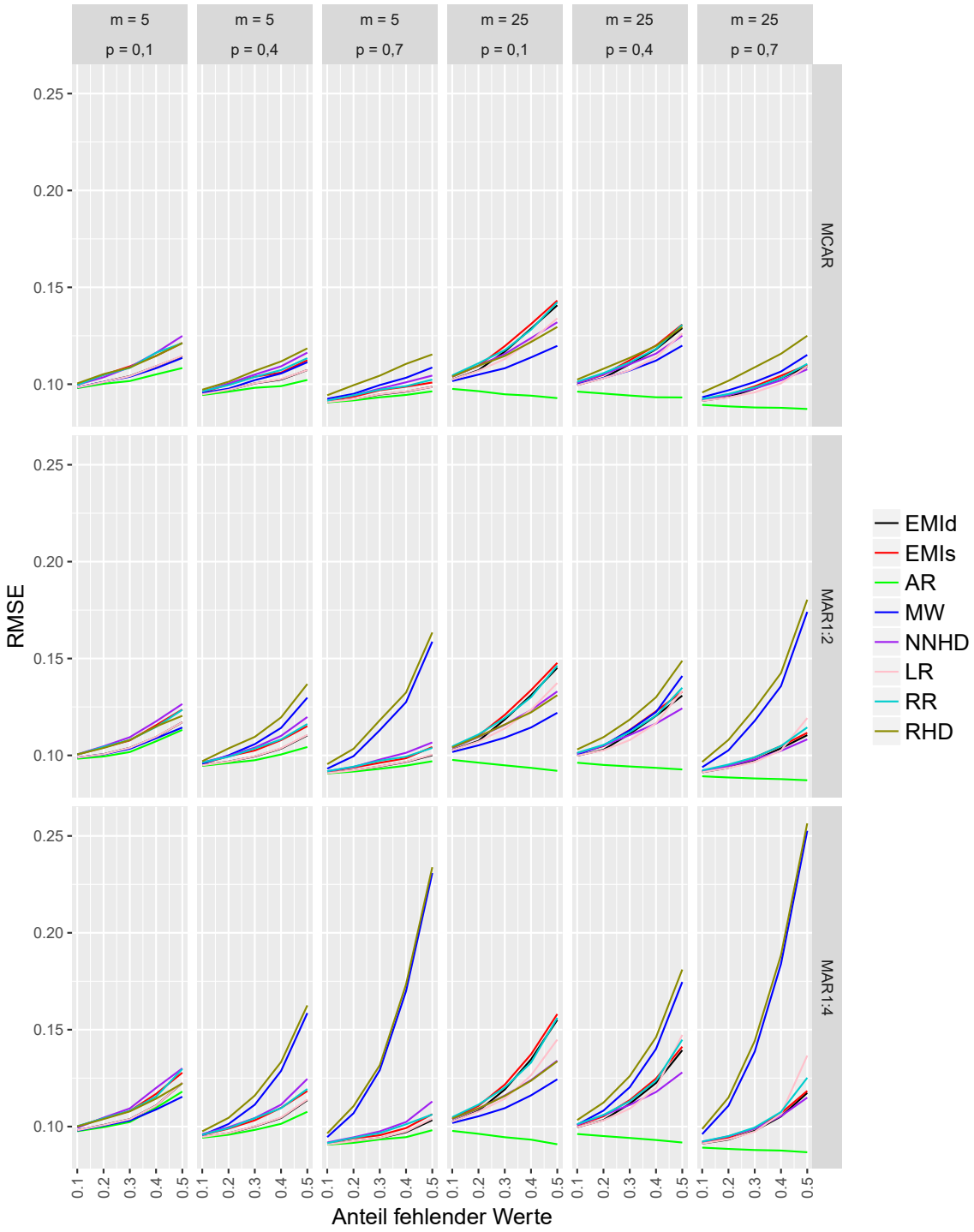

Abbildung 5: RMSE bei $n = 100$ Objekten zwischen wahren und geschätzten Erwartungswerten

nach der Anwendung des jeweiligen Imputationsverfahrens abgetragen ist.

In den kleinsten Datensätzen ($n = 100$, $m = 5$, dargestellt im linken Teil der Abbildung 5) führt die adaptive Regressionsimputation im Mittel zum besten Ergebnis unabhängig von allen anderen Simulationsparametern. Die Mittelwertimputation erreicht bei einer geringen Korrelation häufig ähnlich gute Ergebnisse wie die adaptive Regressionsimputation, verschlechtert sich aber mit zunehmender Korrelation deutlich im Vergleich zu den anderen Imputationsverfahren mit Ausnahme des Random Hot-Decks. Die Ergebnisse der deterministischen EM- und Regressionsimputation sowie die Ergebnisse der stochastischen EM- und Regressionsimputation unterscheiden sich kaum. Dabei sind die Ergebnisse der deterministischen Verfahren immer mindestens genauso gut wie die der stochastischen und häufig besser. Das Nearest-Neighbour Hot-Deck liefert etwas schlechtere Ergebnisse als die stochastischen Versionen der EM- und Regressionsimputation. Das Random Hot-Deck schätzt bei einer Korrelation von $\rho = 0{,}4$ und $\rho = 0{,}7$ die Erwartungswerte am ungenausten und ist auch bei $\rho = 0{,}1$ nur im Mittelfeld der untersuchten Verfahren.

Die Ergebnisse der Datensätze mit $n = 100$ Objekten und $m = 25$ Merkmalen werden im rechten Teil der Abbildung 5 gezeigt. Bei diesen Datensätzen ist die adaptive Regressionsimputation allen anderen Verfahren deutlich überlegen. Sie führt unabhängig vom Ausfallmechanismus und dem Anteil fehlender Werte zur besten Schätzung des Erwartungswerts. Bei einem geringen Anteil fehlender Werte von 10 % oder 20 % sind die Ergebnisse der deterministischen EM- und Regressionsimputation sowie die der stochastischen EM- und Regressionsimputation wieder sehr ähnlich. Bei 30 % und 40 % fehlender Werte verschlechtern sich die Ergebnisse der EM-Methoden zunächst stärker als die der Regressionsimputationen. Bei 50 % fehlender Werte knickt jedoch die deterministische Regressionsimputation noch einmal deutlich stärker ein, wodurch sie bei höheren Korrelationen und den MAR Ausfallmechanismen schlechter wird als die EM-Imputation. Das Nearest-Neighbour Hot-Deck ist im Vergleich zu den deterministischen und stochastischen Regressions- und EM-Imputationsverfahren weniger stark von der Zunahme an fehlenden Werten betroffen, wodurch es bei hohem Anteil fehlender Werte besser als diese vier Verfahren ist. Das Verhalten des Random Hot-Decks und der Mittelwertimputation entspricht bei diesen Datensätzen nahezu dem bei den Datensätzen mit $n = 100$ und $m = 5$.

Bei $n = 500$ Objekten und $m = 5$ Merkmalen ist in der Abbildung 6 häufig kaum ein Unterschied zwischen den Verfahren zu erkennen. Die auffälligste Ausnahme hiervon

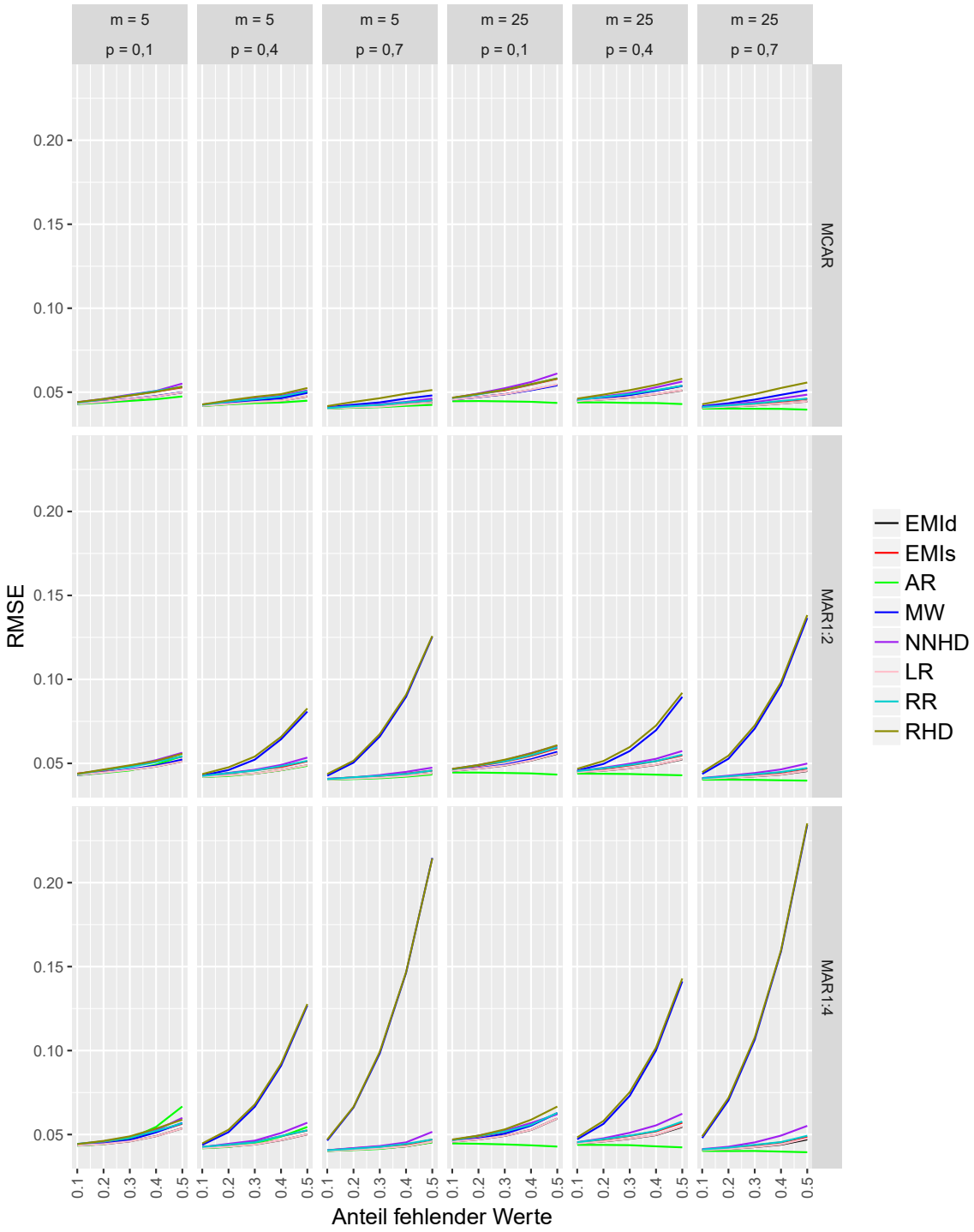

Abbildung 6: RMSE bei $n = 500$ Objekten zwischen wahren und geschätzten Erwartungswerten

bilden das Random Hot-Deck und die Mittelwertimputation, die bei den beiden MAR Ausfallmechanismen deutlich höhere RMSE-Werte als alle anderen Verfahren aufweisen. Eine weitere Ausnahme ist das vergleichsweise schlechte Abschneiden der adaptiven Regressionsimputation bei $\rho = 0{,}1$ bzw. $\rho = 0{,}4$ und einem MAR1:4 Ausfallmechanismus mit vielen fehlenden Werten. Eine genaueren Analyse der Urwerte für die Abbildung 6 zeigt außerdem wieder, dass die deterministischen bzw. stochastischen Versionen der EM- und Regressionsimputation sehr ähnliche Ergebnisse liefern, wobei die deterministischen Methoden leicht besser als die stochastischen sind. Ferner ist das Nearest-Neighbour Hot-Deck meist etwas schlechter als die beiden stochastischen Methoden.

In den größten Datensätzen mit $n = 500$ Objekten und $m = 25$ Merkmalen, dargestellt im rechten Bereich der Abbildung 6, ist die adaptive Regressionsimputation wieder die beste Methode, wobei die Abstände zu den anderen Methoden im Vergleich zu $n = 100$ Objekten geringer sind. Auch bei diesen Datensätzen lassen sich, insbesondere bei Betrachtung der Urwerte, wieder die beiden Zweiergruppen der deterministischen bzw. stochastischen EM- und Regressionsimputation ausmachen. Außerdem sind die deterministischen Verfahren erneut den stochastischen überlegen. Das Nearest-Neighbour Hot-Deck ist meist wieder etwas schlechter als die beiden stochastischen Verfahren. Auch die Ergebnisse des Random Hot-Decks und der Mittelwertimputation, relativ zu den anderen Verfahren betrachtet, entsprechen denen der Datensätze mit $n = 500$ Objekten und $m = 5$ Merkmalen.

Aus den Abbildungen 5 und 6 geht hervor, dass die Schätzgüte bei allen Verfahren mit zunehmendem Anteil fehlender Werte abnimmt. Einzige Ausnahme hiervon scheint die adaptive Regressionsimputation zu sein, deren Ergebnisse bei $m = 25$ Merkmalen unabhängig von der Anzahl fehlender Werte sind bzw. deren Ergebnisse sich teilweise sogar mit zunehmenden Anteil fehlender Werte leicht verbessern. Ferner werden die Verfahren mit Ausnahme der Mittelwertimputation und des Random Hot-Decks mit zunehmender Korrelation und mehr Objekten besser. Der Einfluss der Merkmalsanzahl auf die Ergebnisse ist hingegen wieder nicht eindeutig. So profitiert die adaptive Regressionsimputation vom Übergang von 5 auf 25 Merkmale, während z. B. die Regressions- und EM-Imputationsverfahren bei 100 Objekten sich eher verschlechtern. Unter der Steigerung des Ausfallmechanismus von MCAR hin zu MAR1:4 leiden in besonderem Maße das Random Hot-Deck und die Mittelwertimputation. Der Effekt auf die anderen Verfahren ist deutlich schwächer, jedoch insbesondere bei 100 Objekten durchaus erkennbar.

Über alle Datensätze hinweg betrachtet zeigt sich, dass die deterministische EM- und Regressionsimputation sowie die stochastische EM- und Regressionsimputation häufig zu sehr ähnlichen Ergebnissen führen. Ferner sind die deterministischen Versionen meist den stochastischen Methoden überlegen. Das Nearest-Neighbour Hot-Deck ist mit wenigen Ausnahmen bei den Datensätzen mit $n = 100$ und $m = 25$ sowohl schlechter als die deterministische als auch schlechter als die stochastische EM- und Regressionsimputation. Die Mittelwertimputation und das Random Hot-Deck sind insbesondere bei höheren Korrelationen und stärkerem Ausfallmechanismus gegenüber allen anderen Verfahren deutlich abgeschlagen. Die höchste Schätzgüte weist meist die adaptive Regressionsimputation auf.

3.3 Auswirkungen auf die Varianzschätzung

Die Auswirkungen der Imputationsverfahren auf die Varianzschätzung sind in den Abbildungen 7 und 8 dargestellt. Der Aufbau der Abbildungen 7 und 8 entspricht den vorherigen Abbildungen, nur dass dieses Mal die Abweichung zwischen geschätzter und wahrer Varianz auf der Ordinatenachse abgetragen ist.

In den kleinsten Datensätzen mit $n = 100$ Objekten und $m = 5$ Merkmalen, dargestellt im linken Teil der Abbildung 7, führt die Mittelwertimputation zur schlechtesten Varianzschätzung, gefolgt von der deterministischen EM- und Regressionsimputation, die beide nahezu identische Ergebnisse liefern. Die beste Schätzung resultiert aus der Anwendung der stochastischen EM- oder Regressionsimputation, des Nearest-Neighbour Hot-Decks oder des Random Hot-Decks. Dabei sind die Unterschiede zwischen diesen vier Verfahren nur marginal. Die adaptive Regressionsimputation befindet sich bei einer schwachen und mittleren Korrelation zwischen diesen beiden Gruppen und kann bei $\rho = 0{,}7$ zur Spitzengruppe aufschließen.

Die in der Abbildung 7 gezeigten Ergebnisse der Datensätzen mit $n = 100$ Objekten und $m = 25$ Merkmalen weichen von den Ergebnissen der anderen Datensätzen in einigen Punkten ab. So führen die stochastische EM- und Regressionsimputation häufig zu den zweit- bzw. dritthöchsten RMSE-Werten. Ferner führen ihre deterministischen Pendants im Gegensatz zu allen anderen Datensätzen zu besseren Ergebnissen. Die adaptive Regressionsimputation ist für $\rho = 0{,}1$ und $\rho = 0{,}7$ ähnlich schlecht wie in den anderen Datensätzen, führt aber bei $\rho = 0{,}7$ zur genausten Varianzschätzung. Hingegen bleibt die Mittelwertimputation das ungenauste Verfahren. Auch führen das

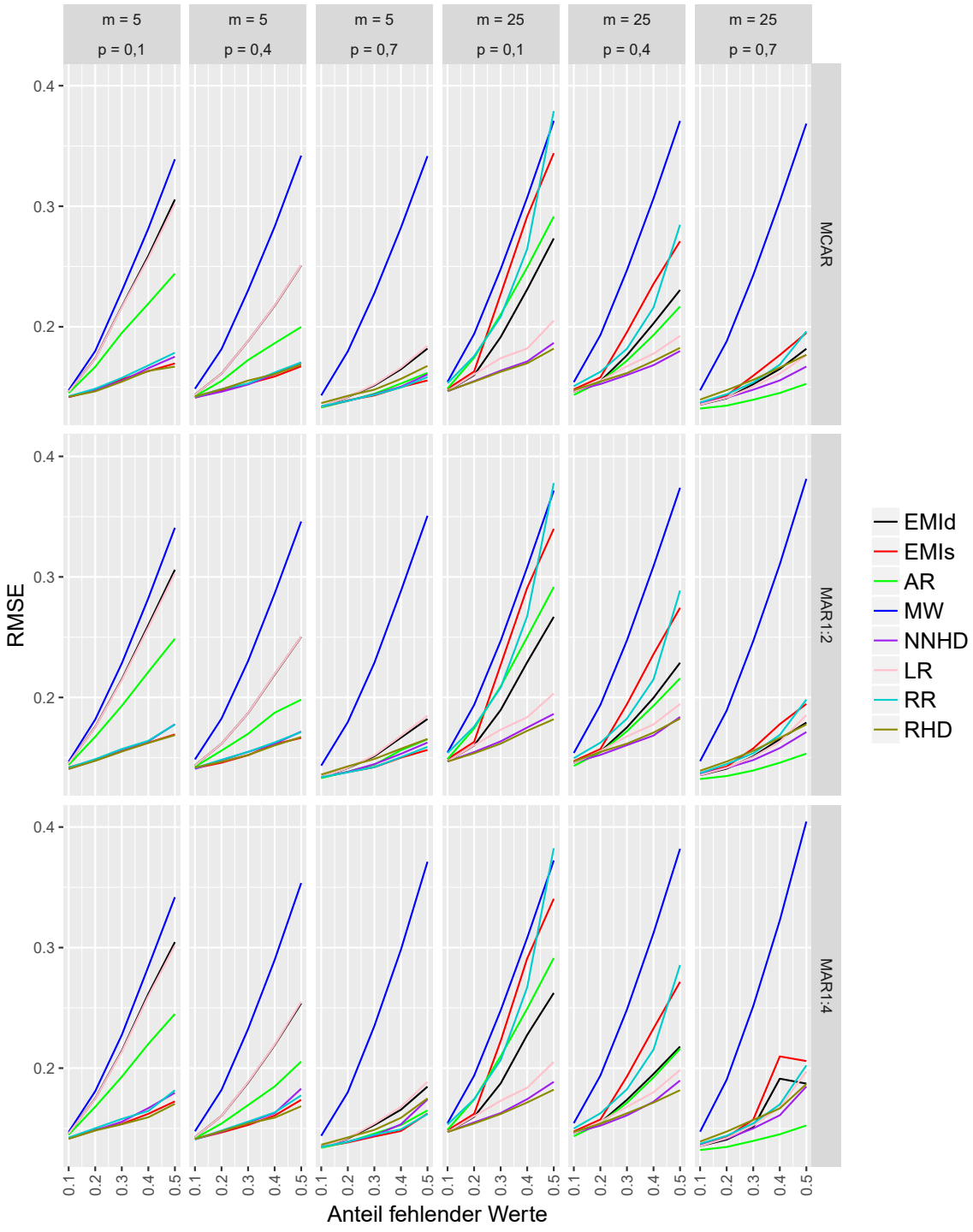

Abbildung 7: RMSE bei $n = 100$ Objekten zwischen wahren und geschätzten Varianzen

Nearest-Neighbour Hot-Deck und das Random Hot-Deck mit Ausnahme der Datensätze mit hoher Korrelation weiterhin zur besten Schätzung der Varianz.

Die Reihenfolge der besten Methoden ist bei den Datensätzen mit $n = 500$ Objekten und $m = 5$ Merkmalen bei $\rho = 0{,}1$ und $\rho = 0{,}4$ (linker Bereich der Abbildung 8) sehr ähnlich zu den Datensätzen mit $n = 100$ Objekten. Es liegt erneut eine Vierteilung der Verfahren in die Gruppe der besten Verfahren (stochastische EM- und Regressionsimputation, Nearest-Neighbour und Random Hot-Deck), der adaptiven Regressionsimputation, der Gruppe bestehend aus der deterministischen EM- und Regressionsimputation sowie der Mittelwertimputation als schlechtestes Verfahren vor. Im Gegensatz zu den Datensätzen mit $n = 100$ Objekten bleibt diese Vierteilung der Verfahren bei $n = 500$ auch bei einer hohen Korrelation bestehen. Beim Übergang von $m = 5$ zu $m = 25$ Merkmalen bei $n = 500$ Objekten ändert sich die Struktur der Ergebnisse nicht wesentlich. Nur die adaptive Regressionsimputation verschlechtert sich relativ zu den anderen Verfahren. Hierdurch fällt sie in die Gruppe der deterministischen Regressions- und EM-Imputation und es resultiert eine Dreiteilung der Verfahren.

Wie bei den vorherigen Kriterien führt auch bei der Varianzschätzung ein höherer Anteil fehlender Werte zu einem schlechteren Ergebnis. Besonders stark betroffen von diesem Effekt ist die Mittelwertimputation und bei geringer sowie mäßiger Korrelation häufig auch die deterministische EM- und Regressionsimputation sowie die adaptive Regressionsimputation. Tendenziell profitieren die meisten Verfahren von einer höheren Objektanzahl und einer stärkeren Korrelation. Hiervon gibt es jedoch einige Ausnahmen, wie z. B. die Ergebnisse der deterministischen Regressionsimputation bei einer schwachen Korrelation und $m = 25$ Merkmalen zeigen. Ferner ist kein Verfahren in der Lage, einen großen Vorteil aus der Erhöhung der Merkmalsanzahl zu ziehen. Vielmehr führt insbesondere bei $n = 100$ Objekten eine Erhöhung der Merkmalsanzahl zu teils deutlich schlechteren Ergebnissen. Die Steigerung des Ausfallmechanismus bewirkt in Verbindung mit einem hohen Anteil fehlender Werte eine Verschlechterung der Ergebnisse. Sie hat bei einem geringen Anteil fehlender Werte jedoch kaum Einfluss. Anders als bei der Schätzung des Erwartungswertes sind alle Verfahren etwa gleichstark von einer Änderung des Ausfallmechanismus betroffen.

Die aus den vorherigen Kriterien bekannten Gruppen der deterministischen bzw. stochastischen EM- und Regressionsimputation ist auch bei der Varianzschätzung wiederzufinden. Im Vergleich zur Erwartungswertschätzung tauschen die deterministischen und stochastischen Verfahren die Rollen, so dass nun die stochastischen Verfahren meist

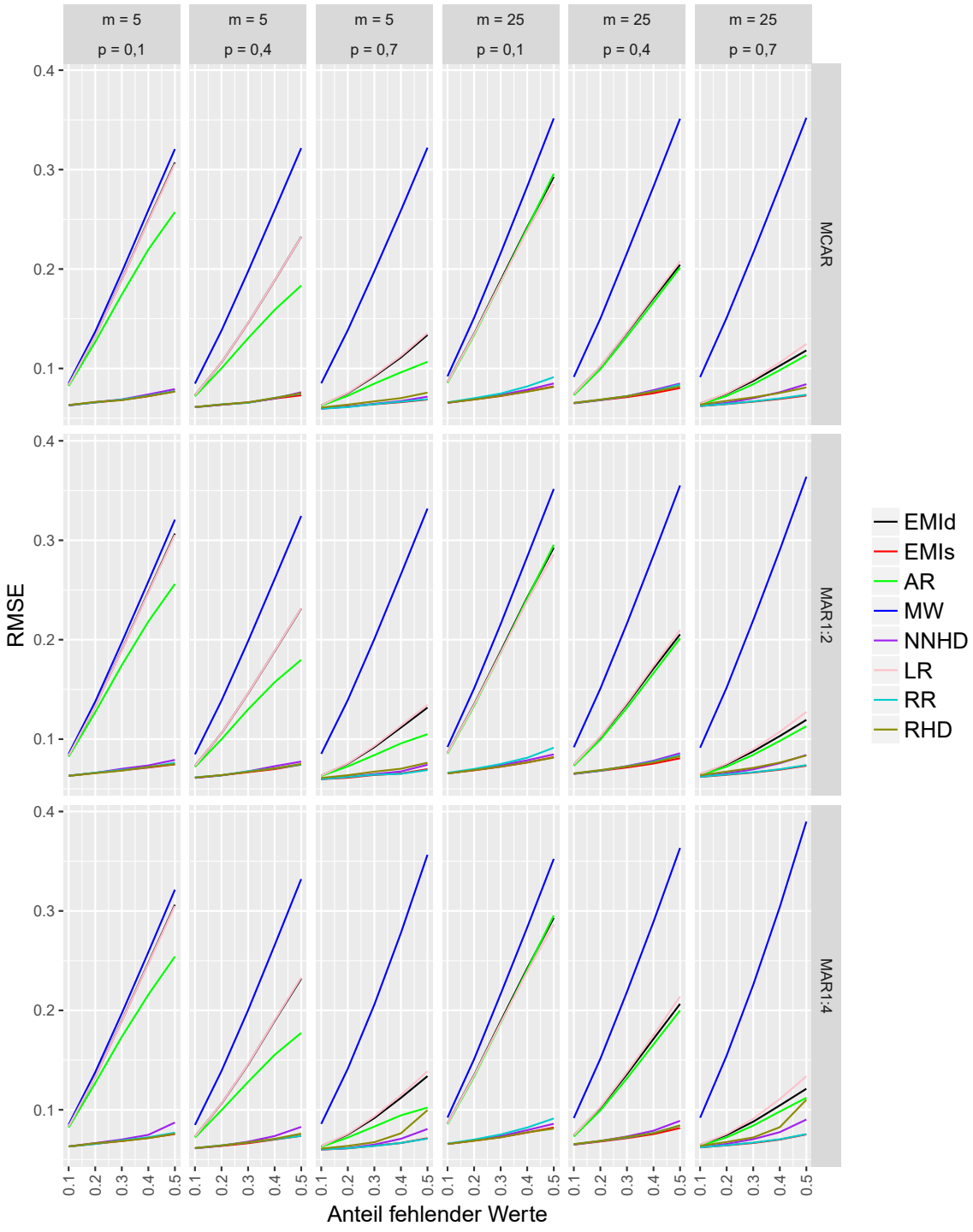

Abbildung 8: RMSE bei $n = 500$ Objekten zwischen wahren und geschätzten Varianzen

zu besseren Ergebnissen führen als die deterministischen Verfahren. Die einzige Ausnahme hiervon bilden die Datensätze mit $n = 100$ Objekten und $m = 25$ Merkmalen. Zur besten Varianzschätzung führen meist das Random Hot-Deck, das Nearest-Neighbour Hot-Deck oder die stochastische EM- oder Regressionsimputation.

3.4 Auswirkungen auf die Korrelationsschätzung

Die Auswirkungen der Imputationsverfahren auf die Korrelationsschätzung sind in den Abbildungen 9 und 10 dargestellt. Der Aufbau der Abbildungen ist analog zu den sechs vorherigen Abbildungen mit der einzigen Ausnahme, dass dieses Mal auf der Ordinatenachse die Abweichungen zwischen der wahren und der geschätzten Korrelation aufgetragen ist.

Die Ergebnisse bei den kleinsten Datensätzen ($n = 100$, $m = 5$, linker Teil der Abbildung 9) und bei den Datensätzen mit $n = 500$ Objekten und $m = 5$ Merkmalen (rechter Teil der Abbildung 10) sind in der Reihenfolge der Verfahren fast identisch. Bei der niedrigen Korrelation von $\rho = 0{,}1$ führen in beiden Fällen die Mittelwertimputation und das Random Hot-Deck zu den geringsten Abweichungen und die adaptive Regressionsimputation zur größten Abweichung. Mit zunehmender Korrelation erhöhen sich jedoch die Abweichungen bei der Mittelwertimputation und dem Random Hot-Deck, so dass sie für $\rho = 0{,}4$ und $\rho = 0{,}7$ zu deutlich größeren Abweichungen führen. Die Ergebnisse des Nearest-Neighbour Hot-Decks und der stochastischen Regressions- und EM-Imputation sind sehr ähnlich. Diese drei Verfahren sind für $\rho = 0{,}4$ und $\rho = 0{,}7$ die besten Verfahren und liegen bei $\rho = 0{,}1$ hinter dem Random Hot-Deck, wobei der Abstand beim größeren Datensatz mit $n = 500$ nur sehr gering ist. Die deterministische EM-Imputation und die deterministische Regressionsimputation führen erneut zu sehr ähnlichen Ergebnissen, die aber leicht schlechter sind als die ihrer stochastischen Gegenstücke. Die adaptive Regressionsimputation ist für $\rho = 0{,}1$ und $\rho = 0{,}4$ schlechter als die beiden deterministischen Verfahren, kann aber für $\rho = 0{,}7$ zu ihnen aufschließen.

Die Reihenfolge der Methoden unterscheidet sich bei $n = 100$ und $m = 25$ erneut von der bei allen anderen Datensätzen. So ist nun die adaptive Regressionsimputation relativ zu den anderen Verfahren besser als in den anderen Datensätzen. Bei $\rho = 0{,}1$ kann sie sich schon gegen die deterministische und stochachstische EM- und Regressionsimputation behaupten. Für $\rho = 0{,}4$ und $\rho = 0{,}7$ führt sie zusammen mit dem Nearest-Neighbour Hot-Deck zu den besten Ergebnissen. Ferner sind im Gegensatz zu

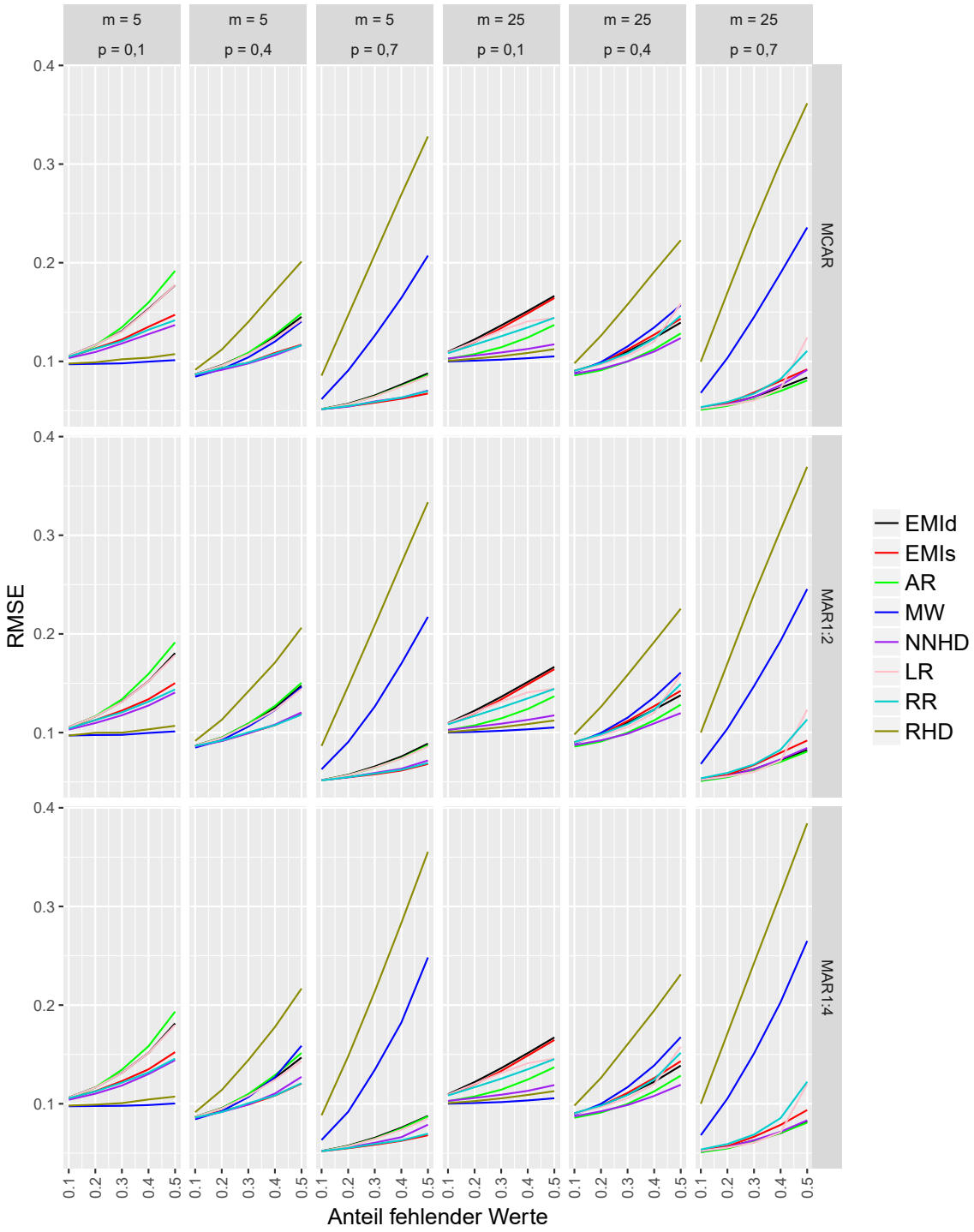

Abbildung 9: RMSE bei $n = 100$ Objekten zwischen wahren und geschätzten Korrelationen

allen anderen Datensätzen bei mittlerer und hoher Korrelation die deterministische EM- und Regressionsimputation ihren stochastischen Gegenparts überlegen. Das Verhalten der Mittelwertimputation und des Random Hot-Decks ähnelt bei den Datensätzen mit $n = 100$ und $m = 25$ dem der anderen Datensätze.

Bei den Datensätzen mit $n = 500$ Objekten und $m = 25$ Merkmalen ergibt sich in der Abbildung 10 ein ähnliches Bild wie bei den vorher beschriebenen Datensätzen mit $m = 5$ Merkmalen. Nur sind jetzt die Ergebnisse der adaptiven Regressionsimputation schon bei $\rho = 0{,}1$ und $\rho = 0{,}4$ sehr ähnlich zu denen der deterministischen Regressions- und EM-Imputation. Eine weitere Abweichung ist, dass die Ergebnisse des Nearest-Neighbour Hot-Decks bei bei $\rho = 0{,}1$ etwas besser und bei $\rho = 0{,}4$ sowie $\rho = 0{,}7$ etwas schlechter als die der stochastischen EM- bzw. Regressionsimputation sind. Die Ergebnisse des Nearest-Neighbour Hot-Decks bleiben aber stets besser als die der deterministischen Regressionsimputation.

Ebenso wie bei den vorherigen Kriterien wird die Korrelationsschätzung bei allen Verfahren mit zunehmenden Anteil fehlender Werte schlechter. Außerdem profitieren die Verfahren erneut vom Übergang von $n = 100$ Objekten zu $n = 500$ Objekten und mit Ausnahme der Mittelwertimputation und des Random Hot-Decks auch wieder von einer höheren Korrelation in den Datensätzen. Hingegen ist erneut kein klarer Trend bei der Verfünffachung der Merkmalsanzahl erkennbar. Mit Ausnahme der Mittelwertimputation und des Random Hot-Decks beeinflusst auch der Ausfallmechanismus die Ergebnisse kaum.

Auch bei der Auswirkung auf die Korrelationsschätzung zeigt sich, dass kaum Unterschiede zwischen der deterministischen Regressions- und EM-Imputation sowie zwischen den beiden stochastischen Varianten bestehen. Ferner ist bei der niedrigen Korrelation von $\rho = 0{,}1$ kein Verfahren in der Lage verlässlich bessere Ergebnisse zu liefern als die Mittelwertimputation und das Random Hot-Deck. Hingegen führt bei $\rho = 0{,}4$ und $\rho = 0{,}7$ die Verwendung des Nearest-Neighbour Hot-Decks oder der stochastischen Regressions- bzw. EM-Imputation zu den geringsten Abweichungen bei der Korrelationsschätzung. Dabei profitiert das Nearest-Neighbour Hot-Deck tendenziell stärker von mehr Merkmalen bei wenigen Objekten, während die beiden stochastischen Methoden eher bei $n = 500$ Objekten dem Nearest-Neighbour Hot-Deck überlegen sind.

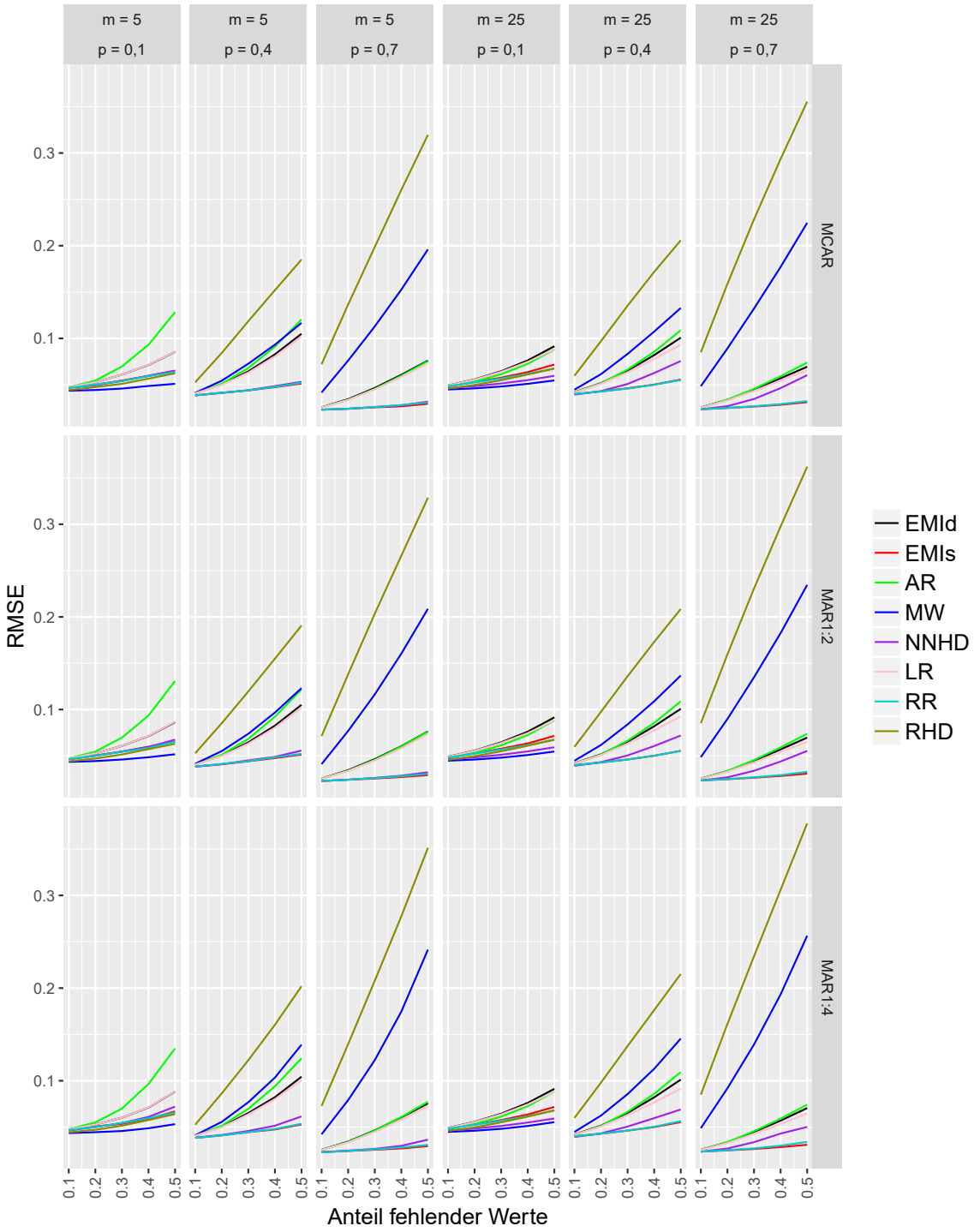

Abbildung 10: RMSE bei $n = 500$ Objekten zwischen wahren und geschätzten Korrelationen

4 Zusammenfassung und Interpretation der Ergebnisse

Beim Vergleich der Ergebnisse über alle Kriterien hinweg zeigt sich, dass es in der Simulation kein Verfahren gibt, dass allen anderen Verfahren in jeder Hinsicht überlegen ist. Es existieren jedoch einige Muster, die über alle Kriterien und/oder Datensätze hinweg auftreten:

In der Simulation liefern sowohl die deterministische EM- und Regressionsimputation als auch die stochastische EM- und Regressionsimputation im Mittel sehr ähnliche Ergebnisse. Bei einem genaueren Vergleich der beiden deterministischen Verfahren zeigt sich, dass einzig bei den Datensätzen mit $n = 100$ Objekten $m = 25$ Merkmalen erkennbare Unterschiede von mehr als 0,01 RMSE in den Abbildungen 3 bis 10 auftreten. In allen anderen Fällen beträgt die Abweichung im Mittel weniger als 0,01 RMSE, wodurch die beiden Verfahren in den Abbildungen de facto nicht zu unterscheiden sind. Diese geringe Abweichung ist (mit der angesprochenen Ausnahme von $n = 100$, $m = 25$) nicht nur im Mittel zu beobachten, sondern sogar bei jedem einzelnen Simulationsdurchlauf, wie aus der Tabelle 1 hervorgeht. In der Tabelle 1 ist die mittlere absolute Abweichung zwischen den beiden deterministischen Verfahren dargestellt, ermittelt anhand der einzelnen Simulationsläufe. Es zeigt sich, dass die mittlere absolute Abweichung zwischen den beiden deterministischen Verfahren mit Ausnahme von $n = 100$, $m = 25$ stets kleiner als 0,005 ist. Bei der Größenordnung, in denen die RMSE-Werte in den Abbildungen liegen, bestätigt dies, dass beide Verfahren nicht nur im Mittel, sondern auch bei jedem Simulationslauf fast identische Ergebnisse liefern. Die größeren Abweichungen bei $n = 100$ und $m = 25$ könnten unter anderem darin begründet sein, dass beide Verfahren bei dem Verhältnis von 4:1 zwischen Objektanzahl und Merkmalsanzahl instabil werden. Dieser Aspekt wird auch durch eine Analyse der Anzahl an benötigten Iterationen des EM-Algorithmus untermauert. Diese Anzahl ist bei den Datensätzen mit $n = 100$ und $m = 25$ deutlich höher als bei allen anderen Datensätzen. Das ist ein Hinweis auf instabile Ergebnisse der EM-Imputation, was zu einer größeren Streuung der Imputationswerte und damit im Endeffekt zu stärkeren Abweichungen führen kann.

Die Unterschiede zwischen der stochastischen EM- und Regressionsimputation sind etwas stärker als bei den beiden deterministischen Versionen. Jedoch sind auch hier alle Punkte mit Ausnahme von Zweien in den Abbildungen 3 bis 10 weniger als 0,03 RMSE-Einheiten auseinander und 90 % der Punkte weniger als 0,0125 RMSE-Einheiten

Objekte	Merkmale	Genauigkeit	Erwartungswert	Varianz	Korrelation
100	5	0,0040	0,0016	0,0028	0,0029
100	25	0,0301	0,0072	0,0248	0,0092
500	5	0,0013	0,0005	0,0011	0,0016
500	25	0,0023	0,0007	0,0027	0,0022

Tabelle 1: Mittlere absolute Abweichung zwischen deterministischer EM- und Regressionsimputation

entfernt. Im Mittel sind sich also beide Verfahren auch sehr ähnlich. Die Werte in der Tabelle 2, die analog zu den Werten in der Tabelle 1 berechnet werden, zeigen, dass die Abweichungen zwischen den beiden stochastischen Verfahren bei den einzelnen Simulationsläufen tendenziell etwas höher als zwischen den deterministischen Verfahren sind. Ein Grund hierfür liegt in der zusätzlichen Varianz der Einzelergebnisse auf Grund der hinzugefügten stochastischen Komponente. Diese Varianz der Einzelergebnisse ist bei den Datensätzen mit $n = 100$ Objekten größer als bei $n = 500$ Objekten. Bei einer Mittelung der Ergebnisse über die 1000 Simulationsläufe verliert diese Varianz der Einzelwert jedoch stark an Einfluss, weshalb auch die beiden stochastischen Verfahren im Mittel sehr ähnlich sind.

Objekte	Merkmale	Genauigkeit	Erwartungswert	Varianz	Korrelation
100	5	0,0368	0,0141	0,0237	0,0134
100	25	0,0250	0,0091	0,0285	0,0090
500	5	0,0163	0,0063	0,0100	0,0061
500	25	0,0107	0,0031	0,0059	0,0022

Tabelle 2: Mittlere absolute Abweichung zwischen stochastischer EM- und Regressionsimputation

Die Auswirkungen der variierten Simulationsparametern auf die Gütekriterien ist in vielen Fällen gleich. So verbessern sich die Ergebnisse aller Verfahren mit zunehmender Objektanzahl. Dieser Effekt ist bei der Betrachtung der Imputationsgenauigkeit am wenigsten stark ausgeprägt, für alle anderen Kriterien aber sehr deutlich erkennbar. Ferner profitieren alle Verfahren mit Ausnahme der Mittelwertimputation und des Random Hot-Decks von einer Erhöhung der Korrelation zwischen den Merkmalen. Die Mittelwertimputation und das Random Hot-Deck profitieren von diesen Faktoren nicht,

da sie als univariate Verfahren die Zusammenhänge zwischen den Merkmalen nicht in die Berechnung der Imputationswerte miteinbeziehen. Der Effekt zusätzlicher Merkmale ist nicht eindeutig. So verbessert sich ein Teil der Verfahren bei zusätzlichen Merkmalen, z. B. die adaptive Regressionsimputation bei fast allen Kriterien. Jedoch verschlechtern sich auch ein Teil der Verfahren, wie z. B. die EM- und Regressionsimputationen beim Übergang von $m = 5$ zu $m = 25$ Merkmalen. Dieser negative Effekt ist bei $n = 100$ Objekten meist deutlich stärker ausgeprägt als bei $n = 500$ Objekten. Ein Grund hierfür könnte das geringe Verhältnis von 4:1 zwischen Objekt- und Merkmalsanzahl sein, wodurch die Schätzung der Imputationsmodellparameter mit einer hohen Unsicherheit behaftet ist.

Für den Ausfall sind folgende Phänomene zu beobachten: Ein höherer Anteil fehlender Werte führt in fast allen Fällen zu einem schlechteren Ergebnis als ein niedrigerer Anteil fehlender Werte. Dieser Effekt tritt besonders deutlich bei den jeweils schlechtesten Verfahren eines Kriteriums auf. Der Einzeleffekt des Ausfallmechanismus ist im Vergleich zum Anteil fehlender Werte schwächer ausgeprägt. Die Ergebnisse verschlechtern sich von MCAR hin zu MAR1:4 bei $\rho = 0{,}1$ meist nur sehr gering, jedoch bei $\rho = 0{,}4$ und $\rho = 0{,}7$ zum Teil deutlich stärker. Es liegen hier also Wechselwirkungen zwischen dem Ausfallmechanismus und der Korrelation im Datensatz vor. Diese Wechselwirkungen sind insbesondere bei der Mittelwertimputation und dem Random Hot-Deck bei der Erwartungswertschätzung sehr gut zu beobachten. Bei einer genaueren Betrachtung der Definition der Ausfallmechanismen sind die gefundenen Wechselwirkungen nicht verwunderlich. Je höher die Korrelation zwischen den Merkmalen ist, desto wahrscheinlicher werden höhere Werte im Merkmal mit fehlenden Werten gelöscht, wodurch der MAR1:2 bzw. MAR1:4 Ausfallmechanismus mit zunehmender Korrelation verstärkt wird.

Ein Vergleich der verschiedenen Gütekriterien zeigt, dass es kein Verfahren gibt, dass in allen Situationen die besten Ergebnisse liefert. Vielmehr ist das beste Verfahren abhängig vom Gütekriterium und von den Eigenschaften des Datensatzes. Die genausten Imputationswerte liefern bei $m = 5$ Merkmalen die deterministische EM- und Regressionsimputation und bei $m = 25$ Merkmalen die adaptive Regressionsimputation. Die adaptive Regressionsimputation führt auch in fast allen Fällen zur besten Schätzung des Erwartungswertes. Bei der Varianzschätzung sind die vier Verfahren Random Hot-Deck, Nearest-Neighbour Hot-Deck sowie stochastische EM- und Regressionsimpuation meist die besten Verfahren mit relativ ähnlichen Ergebnissen. Nur bei $n = 100$ und $m = 25$

Merkmalen ist das Nearest-Neighobur und das Random Hot-Deck den beiden anderen Verfahren deutlich überlegen. Im Gegensatz dazu sind die beiden anderen Verfahren bei $n = 500$ und höheren Korrelationen leicht besser als die Hot-Deck-Verfahren. Die Auswahl des besten Imputationsverfahrens hängt bei der Korrelationsschätzung insbesondere von der Korrelation im Datensatz ab. Bei einer niedrigen Korrelation ($\rho = 0{,}1$) im Datensatz führt die Mittelwertimputation zur besten Schätzung. Bei höheren Korrelationen ist das Nearest-Neighbour Hot-Deck für $n = 100$ Objekte das beste Verfahren und für $n = 500$ Objekte sind es die stochastische EM- bzw. Regressionsimpuation.

Insgesamt zeigt sich, dass bei einer geringen Korrelation von $\rho = 0{,}1$ kein Verfahren sich vom jeweils besten einfachen Verfahren (Mittelwertimputation bei Genauigkeit, Erwartungswerts- und Korrelationsschätzung sowie Random Hot-Deck bei der Varianzschätzung) absetzen kann. Für die effektive Imputation werden also auch bei fortschrittlicheren Imputationsverfahren Merkmale mit einer höheren Korrelation als 0,1 zum imputierenden Merkmal für eine effektive Imputation benötigt. Außerdem erscheint eine Imputation unabhängig von der Zielstellung schwierig, wenn das Verhältnis im Datensatz zwischen Objekt- und Merkmalsanzahl zu klein ist, wie die Datensätze mit $n = 100$ Objekte und $m = 25$ Merkmalen veranschaulichen. Ferner zeigen die Ergebnisse, dass die Beurteilung der Imputationsverfahren vom gewählten Gütekriterium abhängt. Es ist also nicht ohne Weiteres möglich, mit Hilfe der Ergebnisse eines Kriteriums auf die Ergebnisse eines anderen Kriteriums zu schließen. Dies deutet daraufhin, dass bei einer konkreten Untersuchung für die Auswahl eines geeigneten Imputationsverfahrens die spätere Datenanalyse berücksichtigt werden sollte.

5 Fazit

Das Ziel der vorliegenden Simulationsstudie war, verschiedene Imputationsverfahren auf ihre Eignung bei unterschiedlichen Zielstellungen zu untersuchen. Im Rahmen der Simulation kann sich kein Verfahren über alle Kriterien hinweg gegenüber allen anderen Verfahren behaupten. Vielmehr hängt die Eignung eines Verfahrens neben der Zielstellung vom vorliegenden Datensatz ab. Die jeweils besten Verfahren bei einer festen Datensatzgröße sind daher in der Tabelle 3 noch einmal zusammengefasst.

In der Tabelle 3 sind die jeweils besten Imputationsverfahren in Abhängigkeit von der Anzahl der Objekten n, der Anzahl der Merkmale m und dem untersuchten Güte-

n	m	Genauigkeit	Erwartungswert	Varianz	Korrelation
100	5	deterministische Regressions- und EM-Imputation	adaptive Regressionsimputation	stochastische Regressions- und EM-Imputation, Nearest-Neighbour und Random Hot-Deck	Nearest-Neighbour Hot-Deck, stochastische Regressions- und EM-Imputation (Mittelwertimputation, $\rho = 0{,}1$)
100	25	adaptive Regressionsimputation	adaptive Regressionsimputation	Nearest-Neighbour und Random Hot-Deck	Nearest-Neighbour Hot-Deck (Mittelwertimputation, $\rho = 0{,}1$)
500	5	deterministische Regressions- und EM-Imputation	adaptive Regressionsimputation, deterministische Regressions- und EM-Imputation	stochastische Regressions- und EM-Imputation	stochastische Regressions- und EM-Imputation (Mittelwertimputation, $\rho = 0{,}1$)
500	25	adaptive Regressionsimputation	adaptive Regressionsimputation	stochastische Regressions- und EM-Imputation	stochastische Regressions- und EM-Imputation (Mittelwertimputation, $\rho = 0{,}1$)

Tabelle 3: Beste Verfahren bei gegebener Datensatzgröße

kriterium dargestellt. Es zeigt sich, dass die genausten Imputationswerte und die beste Erwartungswertschätzung stets aus der Anwendung einer der drei deterministischen Imputationsverfahren adaptive Regressionsimputation, deterministische Regressions- oder EM-Imputation resultieren. Hingegen führen die stochastischen Gegenparts eher zu einer besseren Schätzung der Varianz bzw. der Korrelation. Eine Sonderstellung nimmt die Mittelwertimputation als deterministisches Verfahren im Rahmen der Korrelationsschätzung ein. Falls eine geringe Korrelation im Datensatz vorliegt, führt sie zur besten Korrelationsschätzung. Falls jedoch höhere Korrelationen vorliegen, verzerrt die Mittelwertimputation die Korrelationsschätzung. Das Nearest-Neighbour Hot-Deck bietet häufig einen guten Kompromiss zwischen der Imputationsgenauigkeit sowie der Erwarwartungswertschätzung und der Varianz- sowie der Korrelationsschätzung.

Als Anwendungsempfehlung resultiert aus der Simulationsstudie, dass für die Auswahl eines Imputationsverfahrens neben dem Untersuchungsziel bzw. der Datenanalysemethode die Datensatzstruktur entscheidend ist. Häufig bieten deterministische Imputationsverfahren genauere Imputationswerte und eine bessere Schätzung des Erwartungswerts, während stochastische Verfahren zur Varianz- bzw. Korrelationsschätzung besser geeignet sind. In der konkreten Anwendung kann die Tabelle 3 als erste Orientierung für die Auswahl eines geeigneten Imputationsverfahrens dienen. Auf Grund der vielfältigen Abhängigkeiten empfiehlt sich jedoch, die Auswahl anhand des vorliegenden Datensatzes und der gewählten Untersuchungsmethode zu prüfen und gegebenenfalls anzupassen.

Literatur

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

Bankhofer, U. (1995): *Unvollständige Daten- und Distanzmatrizen in der Multivariaten Datenanalyse*. Bergisch Gladbach und Köln: Eul.

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

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

Genz, A.; Bretz, F. (2009): *Computation of Multivariate Normal and t Probabilities*. Heidelberg: Springer. DOI: 10.1007/978-3-642-01689-9.

Genz, A.; Bretz, F.; Miwa, T.; Mi, X.; Leisch, F.; Scheipl, F.; Hothorn, T. (2018): *mvtnorm. Multivariate Normal and t Distributions*. R package version 1.0-7.

Kowarik, A.; Templ, M. (2016): Imputation with the R Package VIM. In: *Journal of Statistical Software* 74 (7), S. 1–16. DOI: 10.18637/jss.v074.i07.

Little, R. J. A.; Rubin, D. B. (2002): *Statistical Analysis with Missing Data*. 2. Aufl. Hoboken: Wiley.

Novo, A. A.; Schafer, J. L. (2013): *norm. Analysis of multivariate normal datasets with missing values*. Version 1.0-9.5.

R Core Team (2018): *R. A Language and Environment for Statistical Computing*. Vienna, Austria.

Rockel, T. (2017): *Gütevergleich von Imputationsverfahren. Eine Analyse existierender Simulationsstudien*. Ilmenauer Beiträge zur Wirtschaftsinformatik, Hrsg. von Bankhofer, U.; Nissen, V.; Stelzer, D.; Straßburger, S. Ilmenau: TU Ilmenau.

Schafer, J. L. (1997): *Analysis of Incomplete Multivariate Data*. Boca Raton et al.: Chapman & Hall.

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

van Buuren, S.; Groothuis-Oudshoorn, K. (2011): mice. Multivariate Imputation by Chained Equations in R. In: *Journal of Statistical Software* 45 (3), S. 1–67.