Ilmenauer Beiträge zur Wirtschaftsinformatik

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

Florian Ströhl, Tobias Rockel

Auswirkungen fehlender Daten in der multiplen Regression

Eine Simulationsstudie

Arbeitsbericht Nr. 2020-03, September 2020

Technische Universität Ilmenau
Fakultät für Wirtschaftswissenschaften
Institut für Wirtschaftsinformatik

Autor: Florian Ströhl, Tobias Rockel

Titel: Auswirkungen fehlender Daten in der multiplen Regression – Eine Simulationsstudie

Ilmenauer Beiträge zur Wirtschaftsinformatik Nr. 2020-03, Technische Universität Ilmenau, 2020

ISSN 1861-9223

ISBN 978-3-938940-62-4

urn:nbn:de:gbv:ilm1-2020200403

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

Gliederung

Zusammenfassung: Fehlende Werte stellen in zahlreichen praktischen Anwendungen vielmehr den Regelfall als eine Ausnahme dar, erweisen sich aber bei vielen statistischen Verfahren als störend. Die vorliegende Studie untersucht die Auswirkungen von fehlenden Werten auf die Ergebnisse der multiplen linearen Regression. Dazu werden zunächst spezielle Formen von fehlenden Daten und ausgewählte Verfahren zum Umgang mit diesen vorgestellt. Im Rahmen einer Simulationsstudie werden anschließend die Auswirkungen von verschiedenen Ausfallquoten und -mechanismen anhand von sechs empirischen Datensätzen untersucht. Neben einer Analyse verschiedener Einflussgrößen erfolgt ein Vergleich der vorgestellten Verfahren zur Behandlung der fehlenden Werte. Es zeigt sich, dass keines der untersuchten Verfahren allen anderen Verfahren in jeder Hinsicht überlegen ist und die Wahl des „besten" Verfahrens von der Struktur des Datensatzes und der späteren Verwendung der Regressionsfunktion abhängt. Darüber hinaus konnte festgestellt werden, dass eine Erhöhung der Ausfallquote im Allgemeinen zu einer Verschlechterung der Ergebnisse führt. Die Einflüsse der Objekt- und Merkmalsanzahl hängen von dem jeweiligen Verfahren und den weiteren Eigenschaften des Datensatzes ab und sollten stets zusammen betrachtet werden.

Schlüsselworte: multiple lineare Regression, fehlende Daten, Imputation, Simulationsstudie

1 Einleitung

Die lineare Regression wird bereits seit über einem Jahrhundert verwendet und ist heutzutage Bestandteil jedes Grundlagenbuchs der Statistik. Trotz zahlreicher modernerer Ansätze ist sie nach wie vor ein weitverbreitetes und nützliches Verfahren, das in nahezu allen wissenschaftlichen Disziplinen Anwendung findet (vgl. Fahrmeir et al. 2009, S. 1ff.; James et al. 2013, S. 59). Allerdings setzt die lineare Regression – wie auch viele andere Verfahren aus dem Bereich der Datenanalyse – eine Datenmatrix ohne fehlende Werte voraus (vgl. Schafer und Graham 2002, S. 147). In vielen praktischen Anwendungen stellen fehlende Werte aber vielmehr den Regelfall als eine Ausnahme dar (vgl. Backhaus und Blechschmidt 2009, S. 266; Enders 2010, S. 1). So zeigte zum Beispiel eine Untersuchung von Eekhout et al. (2012), dass von 285 betrachteten epidemiologischen Studien, 262 Studien Datensätze mit fehlenden Werten enthielten und es bei weiteren 19 Studien unklar war, ob fehlende Werte aufgetreten sind. In lediglich 4 von 285 Studien wurden fehlende Werte explizit durch die Autoren ausgeschlossen. Bei zahlreichen empirischen Untersuchungen ist somit zunächst eine Behandlung der fehlenden Werte notwendig, bevor die lineare Regression durchgeführt werden kann. Dabei ist zu beachten, dass eine unbedachte Behandlung der fehlenden Werte zu einer Verzerrung der Ergebnisse führen kann (vgl. Rockel 2018, S. 1).

Das Ziel dieses Arbeitspapiers besteht darin, die Auswirkungen von fehlenden Daten auf die Ergebnisse der multiplen linearen Regression zu untersuchen. Dazu werden neben verschiedenen Ausfallmechanismen und Verfahren zum Umgang mit fehlenden Daten, auch deren Auswirkungen auf die Ergebnisse der Regressionsfunktion anhand von sechs verschiedenen Datensätzen betrachtet. Dabei wird entlang der Forschungsfrage „Wie wirken sich fehlende Daten auf die Ergebnisse der multiplen linearen Regression aus?“ eine Simulationsstudie durchgeführt, um zu ermitteln, unter welchen Bedingungen fehlende Daten zu einer ähnlichen bzw. einer sehr verschiedenen Regressionsfunktion im Vergleich zu vollständigen Daten führen. Neben einer Analyse verschiedener Einflussgrößen erfolgt auch ein Vergleich von Verfahren zur Behandlung von fehlenden Daten. Des Weiteren soll auch die folgende Frage beantwortet werden: „Welchen Einfluss hat die Objektanzahl eines Datensatzes auf die Auswirkungen fehlender Daten im Rahmen der multiplen linearen Regression?“

Im Kapitel 2 werden zunächst die theoretischen Grundlagen für die Simulation gelegt. Anschließend werden in Kapitel 3 die verwendeten Datensätze und der allgemeine Aufbau der

Simulationsstudie beschrieben, bevor in Kapitel 4 die Ergebnisse der Simulationsstudie dargestellt werden. Diese Ergebnisse werden in Kapitel 5 ausgewertet und interpretiert. Den Abschluss dieses Arbeitspapiers bilden eine Zusammenfassung der gewonnenen Erkenntnisse sowie ein Ausblick auf zukünftige Forschungsfragen.

2 Theoretische Grundlagen

2.1 Multiple lineare Regression

Regressionsverfahren verfolgen das Ziel, ausgehend von einer oder mehreren unabhängigen Variablen (auch als Regressoren oder Kovariablen bezeichnet) eine abhängige Variable (auch als Regressand oder Zielvariable bezeichnet) zu beschreiben. Eine Unterscheidung der Regressionsverfahren kann nach dem Skalenniveau der abhängigen und unabhängigen Variablen erfolgen (vgl. Fahrmeir et al. 2009, S. 19).[1] Dabei werden im Rahmen der multiplen linearen Regression zunächst lediglich quantitative Variablen berücksichtigt:

$$y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \cdots + \beta_k * x_k + \varepsilon \tag{1}$$

Dieses Modell unterstellt einen linearen Zusammenhang zwischen der abhängigen Variable y und den unabhängigen Variablen $x_1, \dots, x_k$, der durch die Regressionskoeffizienten $\beta_1, \dots, \beta_k$ beschrieben wird und durch eine Störgröße ε additiv überlagert wird. Neben quantitativen Merkmalen können auch qualitative Merkmale mittels Dummy-Codierung in ein solches Modell einfließen (vgl. James et al. 2013, S. 84). Details zur Parameterschätzung sind zum Beispiel bei Bankhofer und Vogel (2008, 227ff.) zu finden.

Die Güte eines Regressionsmodells kann unter anderem mit Hilfe des korrigierten Bestimmtheitsmaßes $\bar{R}^2$ beurteilt werden, welches auf dem (unkorrigierten) Bestimmtheitsmaß R^2 basiert (vgl. Fahrmeir et al. 2009, S. 98 und S. 160f.):

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}\hat{\varepsilon}_i{}^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2}$$

$$\bar{R}^2 = 1 - \left(\frac{n-1}{n-k-1}\right) * (1 - R^2) \tag{3}$$

[1] Für einen Überblick verschiedener Regressionsverfahren sei der interessierte Leser z. B. auf Fahrmeir et al. (2009, S. 55ff.) verwiesen.

Neben der Beschreibung des Zusammenhanges zwischen der abhängigen und den unabhängigen Variablen wird auch eine Prognose der abhängigen Variable anhand der geschätzten Regressionskoeffizienten ermöglicht (vgl. Bankhofer und Vogel 2008, 228ff.).

2.2 Ausfallmechanismen und der Umgang mit fehlenden Werten

Die Definition von Little und Rubin (2020, S. 4) „Missing data are unobserved values that would be meaningful for analysis if observed; in other words, a missing value hides a meaningful value“ zeigt die Bedeutung im Umgang mit fehlenden Werten. Der unreflektierte Umgang mit fehlenden Werten kann zu einer Verzerrung der Untersuchungsergebnisse führen. Dabei spielt der dem Fehlen der Daten zugrundeliegende Ausfallmechanismus eine wesentliche Rolle in der adäquaten Behandlung der fehlenden Daten (vgl. Rockel 2018, S. 1). Der Ausfallmechanismus gibt Auskunft darüber, welcher Zusammenhang zwischen der Ausfallwahrscheinlichkeit und den Ausprägungen der Merkmale besteht. Die Unterscheidung in drei verschiedene Ausfallmechanismen geht auf Rubin (1976) zurück.

Die Daten werden als **Missing At Random (MAR)** bezeichnet, wenn die Ausfallwahrscheinlichkeit nicht von den fehlenden Werten abhängt, aber von den beobachteten Werten abhängen kann. Infolgedessen hängt die Verteilung der fehlenden Werte bzw. die Ausfallwahrscheinlichkeit nur von den beobachteten Werten ab (vgl. Bankhofer 1995, S. 13; Enders 2010, S. 6 und S. 11; Little und Rubin 2020, S. 14). Als **Missing Completely At Random (MCAR)** werden die Daten bezeichnet, bei denen die Ausfallwahrscheinlichkeit weder von den fehlenden Werten noch von den beobachteten Werten abhängt. Der MCAR Ausfallmechanismus stellt somit einen (restriktiveren) Spezialfall des MAR Ausfallmechanismus dar, bei dem das Fehlen der Werte in keiner Relation zu den beobachteten und fehlenden Werten der Datenmatrix steht (vgl. Bankhofer 1995, S. 14 – 16 und S. 21f.; Enders 2010, S. 7 und S. 12; Little und Rubin 2020, S. 13f.). Die Daten werden als **Missing Not At Random (MNAR)** bezeichnet, wenn die Ausfallwahrscheinlichkeit von den fehlenden Werten selbst abhängt, und infolgedessen die Bedingungen des MAR Ausfallmechanismus nicht erfüllt sind. Zudem kann die Verteilung der fehlenden Werte bzw. die Ausfallwahrscheinlichkeit auch von den beobachteten Werten abhängen (vgl. Enders 2010, S. 8 und S. 11; Little und Rubin 2020, S. 14). Der MNAR Ausfallmechanismus kann als stärkste Form des Ausfalls angesehen werden (vgl. Rockel 2017, S. 25).

Zur Behandlung fehlender Werte können in Anlehnung an Bankhofer (1995, S. 3f. und S. 89ff.) die folgenden fünf Verfahrenskategorien unterschieden werden:

- Eliminierungsverfahren
- Imputationsverfahren
- Parameterschätzverfahren
- Multivariate Analyseverfahren
- Sensitivitätsbetrachtungen

Im Rahmen dieses Arbeitspapiers sollen verschiedene Eliminierungs- und Imputationsverfahren näher untersucht werden. Für die Betrachtung anderer Verfahren zum Umgang mit fehlenden Daten sei der interessierte Leser z. B. auf Bankhofer (1995, S. 89ff.) sowie Little und Rubin (2020, S. 23ff.) verwiesen.

Eliminierungsverfahren sind auch in Zeiten modernerer Verfahren weit verbreitete, wenn nicht sogar die meist genutzten, Verfahren zur Behandlung von fehlenden Daten (vgl. Bartlett et al. 2015, S. 730; Demissie et al. 2003, S. 546; Enders 2010, S. 37ff.; Peugh und Enders 2004, S. 536ff.; Schafer und Graham 2002, S. 155). Dies beruht unter anderem darauf, dass Eliminierungsverfahren in vielen Statistik-Softwarepaketen vorinstalliert und in zahlreichen Implementierungen als Default-Option hinterlegt sind (vgl. Bartlett et al. 2014, S. 720; Enders 2010, S. 39; Glynn und Laird 1986, S. 3; Little 1992, S. 1229; Schafer und Graham 2002, S. 155).

Die Behandlung der fehlenden Daten erfolgt, indem die Objekte oder Merkmale mit fehlenden Werten von der weiteren Untersuchung ausgeschlossen werden. Dabei werden im Rahmen der Objekteliminierung alle Objekte, die in mindestens einem Merkmal einen fehlenden Wert aufweisen, ausgeschlossen. Bei der Merkmalseliminierung erfolgt ein Ausschluss aller Merkmale, die bei mindestens einem Objekt nicht beobachtet wurden. In beiden Fällen resultiert eine vollständige Datenmatrix, die mittels herkömmlicher Analyseverfahren untersucht werden kann. Da die Objekteliminierung das geeignetere Vorgehen zur Untersuchung der Struktur der Merkmale dargestellt, soll die Merkmalseliminierung im Rahmen dieser Studie nicht betrachtet werden. Die Analyse der vollständig erhobenen Objekte (bzw. die Eliminierung aller Objekte mit fehlenden Werten) wird auch als complete-case analysis (CCA) oder listwise deletion (LD) bezeichnet (vgl. Bankhofer 1995, S. 91 und S. 98; Frane 1976, S. 409).

Die Behandlung der fehlenden Werte im Rahmen von Imputationsverfahren erfolgt, indem die vorhandenen Werte um Schätzungen für die fehlenden Werte ergänzt werden. Da die fehlenden Werte gewissermaßen durch die geschätzten Werte ersetzt werden, wird in der Literatur auch die Bezeichnung Ersetzungsverfahren verwendet (vgl. Bankhofer 1995, S. 104f.; Enders 2010, S. 42). In dieser Studie wird sich dabei auf Imputationsverfahren beschränkt, die für jeden fehlenden Wert genau einen Imputationswert bestimmen (engl. single imputation methods). Für die Betrachtung der multiplen Imputation, d. h. Imputationsverfahren bei denen mehrere Imputationswerte für jeden fehlenden Wert bestimmt werden, sei der interessierte Leser z. B. auf Little und Rubin (2020, S. 95ff.) oder van Buuren (2018, S. 29ff.) verwiesen.

Eines der einfachsten Imputationsverfahren stellt die Ersetzung aller fehlenden Werte in einem Merkmal durch den arithmetischen Mittelwert der beobachteten Werte im selben Merkmal dar (vgl. Bankhofer 1995, S. 106f.; Enders 2010, S. 42). Neben der Imputation eines Mittelwerts (IMW) ist auch die Imputation durch Regression ein weitverbreitetes Verfahren zum Umgang mit fehlenden Werten, welches nach Little und Rubin (2020, S. 70) eine Verbesserung gegenüber der IMW darstellt. Im Unterschied zur IMW wird zur Imputation der fehlenden (quantitativen) Werte eines Merkmals kein Lageparameter, sondern die mittels eines Regressionsmodells geschätzten Werte, verwendet. In den meisten Anwendungen wird dabei ein linearer Zusammenhang zwischen den Merkmalen unterstellt, so dass ein multiples lineares Regressionsmodell zur Schätzung der fehlenden Werte genutzt wird (vgl. Bankhofer 1995, S. 126; Enders 2010, S. 44). Falls die Prognosewerte eines Modells direkt als Imputationswerte verwendet werden, bezeichnet man dieses Verfahren als **deterministische Regressionsimputation (DRI)**. Wenn die prognostizieren Werte noch mit einer zusätzlichen Störgröße versehen werden, wird das Verfahren als **stochastische Regressionsimputation (SRI)** bezeichnet.

Eine weitere Klasse an Imputationsverfahren bilden die sogenannten Hot-Deck Imputationsverfahren. Diese Verfahren zeichnen sich dadurch aus, dass die fehlenden Merkmalsausprägungen eines Objektes ausschließlich durch die beobachteten Merkmalsausprägungen eines anderen, ähnlichen Objektes (bzw. mehrerer, ähnlicher Objekte) aus derselben Datenmatrix ersetzt werden. Die Objekte, die fehlende Werte aufweisen, werden dabei als Empfänger (engl. recipient) und die Objekte, die die Imputationswerte liefern als Spender (engl. donor) bezeichnet. Im Rahmen der Hot-Deck Imputationsverfahren erfolgt somit keine Aggregation

mehrerer vorhandener Werte zur Bestimmung der Imputationswerte, so dass diese Verfahren durch eine Verdopplung bereits vorhandener Werte charakterisiert sind. Je nach gewählten Hot-Deck Verfahren ist die Aufteilung der Objekte in Spender und Empfänger nicht zwangsläufig disjunkt, sodass ein Objekt gleichzeitig Empfänger und Spender sein kann (vgl. Bankhofer 1995, S. 120; Ford 1983, S. 186; Joenssen 2015b, S. 56 und S. 61ff.; Sande 1983, S. 341).

Im Rahmen der Studie werden zwei Hot-Deck Imputationsverfahren betrachtet. Bei der **Simple Random Hot-Deck Imputation (SRHDI)** werden die Spender durch eine Zufallsauswahl bestimmt. Dies stellt nach Enders (2010, S. 49) die einfachste Version einer Hot-Deck Imputation dar. Dabei sollen im Rahmen der Studie die fehlenden Merkmale eines Empfängers sequenziell durch die Ausprägungen mehrerer Objekte (Spender) ersetzt werden. Diese Art der Imputation wird von Bankhofer (1995, S. 109ff.) zu den Imputationen mittels Zufallsauswahl gezählt. Neben einer einfachen Anwendung besitzt dieses Vorgehen den Vorteil, dass für jedes Merkmal die maximal mögliche Menge an Spendern zur Verfügung steht (vgl. Joenssen 2015b, S. 99; Sande 1983, S. 342).

Allerdings kann die zufällige Auswahl eines Spenders zu verzerrten Ergebnissen führen, wenn kein MCAR Ausfallmechanismus vorliegt. Des Weiteren scheint die Verwendung eines ähnlichen (einzelnen) Spenders zweckmäßiger, um nachfolgende Analysen durchzuführen, da so die Schätzungen verbessert werden können (vgl. Andridge und Little 2010, S. 49; Bankhofer 1995, S. 120; Joenssen 2015b, S. 68 und S. 92; Joenssen und Bankhofer 2012, S. 64). Erfolgt die Auswahl des ähnlichsten Spenders mit Hilfe eines Distanzmaßes, wird in der Literatur die Bezeichnung **Nearest Neighbor Hot-Deck Imputation (NNHDI)** verwendet. Als Spender wird anschließend das Objekt ausgewählt, das die geringste Distanz zum Empfänger aufweist (vgl. Andridge und Little 2010, S. 44; Joenssen 2015b, S. 78; Schnell 1986, S. 111f.).

3 Design der Simulationsstudie

Die Untersuchung der Auswirkungen von fehlenden Daten auf die Ergebnisse der multiplen linearen Regression erfolgt im Rahmen des Arbeitspapiers mittels einer Simulationsstudie. Dazu werden vollständige reale Datensätze und künstlich erzeugte Ausfallmechanismen verwendet. Neben einer Analyse der Auswirkungen verschiedener Einflussgrößen, soll ein Vergleich der thematisierten MD-Verfahren erfolgen, um Anwendungsempfehlungen ableiten

zu können. Zur Durchführung der Simulationsstudie wird die Statistikprogrammiersprache R (R Core Team 2020) in der Version 3.6.3 und das Anwendungsprogramm RStudio (RStudio Team 2020) in der Version 1.2.5033 verwendet. Um die Auswirkungen fehlender Daten auf die Ergebnisse der multiplen linearen Regression zu untersuchen, werden sechs Datensätze betrachtet. Die Auswahl dieser sechs Datensätze ermöglicht es, die Auswirkungen fehlender Daten anhand von zwei Größenordnungen der Merkmalsanzahl und unterschiedlichen Objektanzahlen zu untersuchen. Im Folgenden soll eine kurze Beschreibung der Datensätze erfolgen.

Der **Autodatensatz** weist insgesamt 9 Merkmale und 392 vollständige Objekte auf. Im Rahmen der multiplen linearen Regression soll der *Verbrauch* eines Fahrzeuges anhand von sechs quantitativen Merkmalen, wie z. B. *Anzahl der Zylinder* und *Gewicht* des Fahrzeuges und dem nominalen Merkmal *Herkunftsland* beschrieben werden.[2] Das Merkmal *Herkunftsland* wird über eine Dummy-Variable modelliert, indem zwischen amerikanischen und nicht amerikanischen Fahrzeugen unterschieden wird. Darüber hinaus enthält dieser Datensatz ein weiteres nominales Merkmal (*Fahrzeugname*). Da die Verwendung von 391 Dummy-Variablen zur Berücksichtigung dieses Merkmals in der multiplen linearen Regression wenig zweckmäßig erscheint, wird dieses Merkmal in der Studie nicht berücksichtigt werden.

Der **Irismuschelndatensatz** weist insgesamt 9 Merkmale und 4177 vollständige Objekte auf. Zur Prognose der *Anzahl an Zuwachsringen*, werden 8 Merkmale, wie z. B. das *Geschlecht* und die *Länge* der Muscheln, verwendet.[3] Das kategoriale Merkmal Geschlecht wird über zwei Dummy-Variablen modelliert, da es neben den Ausprägungen „Female“ und „Male“ noch die Ausprägung „Infant“ aufweist. Die Ausprägung „Infant“ wird als eigenes Geschlecht aufgeführt, weil das Geschlecht einer Irismuschel nicht bei der Geburt bestimmt wird (vgl. Mehta 2019, S. 43).

[2] Der verwendete Autodatensatz kann über die folgende URL heruntergeladen werden: https://www.kaggle.com/uciml/autompg-dataset (Zugriff: 02.12.2019). Sechs weitere Objekte, bei denen die Ausprägung des Merkmals *horsepower* fehlt, wurden für die weitere Untersuchung entfernt.

[3] Der verwendete Irismuscheldatensatz kann über die folgenden URLs heruntergeladen werden: https://www.kaggle.com/rodolfomendes/abalone-dataset (Zugriff: 02.12.2019) oder https://archive.ics.uci.edu/ml/datasets/abalone (Zugriff: 02.12.2019). Dieser Datensatz enthält keine fehlenden Werte mehr, da die Objekte mit fehlenden Werten bereits entfernt wurden. Des Weiteren wurden die quantitativen Merkmale: *Length*, *Diameter*, …, *Shell.weight* um den Faktor 200 skaliert, um die Daten für ein künstliches neurales Netz zu nutzen. Für die Studie wurden die Daten wieder auf ihre ursprünglichen Werte zurückskaliert.

Der **Kalifornien-Hauspreisdatensatz** weist insgesamt 10 Merkmale und 20433 vollständige Objekte auf. Zur Prognose der *Median-Hauspreise* von kalifornischen Blockgruppen werden acht quantitative Merkmale, wie z. B. das *Median-Einkommen* und die Anzahl der *Räume in jeder Blockgruppe*, verwendet. Eine Blockgruppe stellt dabei die kleinste geografische Einheit dar, für die das US Census Bureau Beispieldaten veröffentlicht. Das nominale Merkmal *Ozeannähe* soll im Rahmen der Studie nicht betrachtet werden, da dieses in der erstmaligen Verwendung dieses Datensatzes von Pace und Barry (1997) nicht enthalten war und erst nachträglich zum Datensatz hinzugefügt wurde.[4]

Der **Hittersdatensatz** weist insgesamt 20 Merkmale und 263 vollständige Objekte auf. Zur Prognose des *Jahresgehalts* von Baseballspielern werden 19 Merkmale, wie z. B. die Anzahl der *Homeruns des Vorjahres* und die *Gesamttrefferanzahl* eines Spielers, verwendet. Dabei werden drei (binäre) nominale Merkmale über jeweils eine eigene Dummy-Variable modelliert.[5] Anhand dieses Datensatzes können insbesondere die Auswirkungen fehlender Daten bei einem relativ geringem Verhältnis zwischen Objekt- und Merkmalsanzahl bzw. der Anzahl zu schätzender Regressionskoeffizienten untersucht werden.

Der **Krebsdatensatz** weist insgesamt 33 Merkmale und 3047 Objekte auf. Zur Prognose der *mittleren Krebssterblichkeit* stehen bis zu 32 Merkmale, wie z. B. der *Median des Alters der weiblichen* bzw. *männlichen Bevölkerung* eines Landes, zur Verfügung. Allerdings liegen nur 591 vollständige Objekte vor. Da lediglich drei Merkmale fehlende Werte aufweisen, sollen diese Merkmale von der weiteren Untersuchung ausgeschlossen werden. Außerdem werden 12 weitere Merkmale von der Untersuchung ausgeschlossen, da für einige Merkmale, wie z. B. das nominale Merkmale *Land*, eine große Anzahl Dummy-Variablen erforderlich wäre. Durch dieses Vorgehen wird zudem eine ähnliche Merkmalsanzahl wie bei dem Hittersdatensatz erreicht.[6]

---

[4] Der verwendete Kalifornien-Hauspreisdatensatz kann über die folgende URL heruntergeladen werden: https://www.kaggle.com/harrywang/housing (Zugriff: 05.02.2020). 207 weitere Objekte, bei denen die Ausprägung des Merkmals *Anzahl der Schlafzimmer* fehlt, wurden für die weitere Untersuchung entfernt.

[5] Der verwendete Hittersdatensatz kann über die folgende URL heruntergeladen werden: https://www.kaggle.com/floser/hitters (Zugriff: 02.12.2019). 59 weitere Objekte, bei denen die Ausprägung für die abhängige Variable *Jahresgehalt* fehlt, wurden für die weitere Untersuchung entfernt.

[6] Der verwendete Krebsdatensatz kann über die folgende URL heruntergeladen werden: https://data.world/exercises/linear-regression-exercise-1 (Zugriff: 06.02.2020). Die für die weitere Untersuchung entfernten Merkmale sind: *binnedinc*, *geography*, *pctnohs18_24*, *pcths18_24*, *pctsomecol18_24*, *pctbachdeg18_24*, *pcths25_over, pctbachdeg25_over*, *pctemployed16_over*, *pctunemployed16_over*, *pctprivatecoveragealone*, *pctwhite*, *pctblack*, *pctasian* und *pctotherrace*.

Der **Energiedatensatz** weist insgesamt 29 Merkmale und 19735 vollständige Objekte auf. Zur Prognose des *Gesamtenergieverbrauchs* eines Hauses stehen somit bis zu 28 Merkmale, wie z. B. die *Temperatur* und *Luftfeuchtigkeit des Wohnzimmers*, zur Verfügung. Für die weitere Untersuchung werden das Merkmal *Datum* und neun weitere Merkmale entfernt. Durch dieses Vorgehen wird eine ähnliche Merkmalsanzahl wie bei dem Hitters- und Krebsdatensatz erreicht. Neben den beiden Merkmalen *rv1* und *rv2*, die jeweils über eine Zufallsvariable erzeugt wurden, wurden sieben weitere als nicht signifikant ausgewiesene Merkmale entfernt. Das Bestimmtheitsmaß R^2 verringert sich durch die Entfernung dieser neun Merkmale um lediglich $0{,}005$.[7]

Tabelle 1 zeigt eine Zusammenfassung der Datensätze und gibt eine Übersicht zu den resultierenden multiplen linearen Regressionsmodellen. Neben der Anzahl der Objekte n und unabhängigen Variablen k sind auch die (korrigierten) Bestimmtheitsmaße der jeweiligen Regressionsmodelle angegeben.

Datensatz	**Abhängige Variable y**	**Objektanzahl n**	**Anzahl unabhängiger Variablen k**	**R^2**	**$\bar{R}^2$**
Auto	*Verbrauch*	392	7	0,8241	0,8209
Irismuschel	*Rings*	4177	9	0,5379	0,5369
Kalifornien-Hauspreis	*Median-Hauspreis*	20433	8	0,6369	0,6368
Hitters	*Jahresgehalt*	263	19	0,5461	0,5106
Krebs	*mittlere Krebssterblichkeit*	3047	17	0,4528	0,4498
Energie	*Energieverbrauch*	19735	18	0,1644	0,1636

Tabelle 1: *Datensätze und Regressionsmodelle im Vergleich (Quelle: Eigene Darstellung)*

In einigen Datensätzen scheint die Unterstellung eines linearen Zusammenhanges wenig zweckmäßig, da trotz der teilweise recht hohen Anzahl unabhängiger Variablen nur ein geringer Teil der Varianz der Ausgangsdaten erklärt werden kann. Eine alternative Modellwahl

[7] Der verwendete Energiedatensatz kann über die folgende URL heruntergeladen werden: https://www.kaggle.com/loveall/appliances-energy-prediction (Zugriff: 25.02.2020). Die für die weitere Untersuchung entfernten Merkmale sind: *date, rv1, rv2, T1, RH_4, RH_5, T5, T7, RH_9* und *Press_mm_hg*. Die Verringerung des Bestimmtheitsmaßes in Höhe von $0{,}005$ berechnet sich wie folgend dargestellt: $0{,}1649 - 0{,}1644 = 0{,}005$.

zur Analyse dieser Datensätze soll im Rahmen der Studie nicht untersucht werden, da das Ziel die Untersuchung der Auswirkungen fehlender Daten und nicht die bestmögliche Analyse dieser sechs Datensätze ist.

Zur Erzeugung der fehlenden Daten werden ein MCAR, ein MAR und ein MNAR Ausfallmechanismus genutzt. Dabei wird ein allgemeines Ausfallmuster in den ersten $\frac{m}{2}$ Merkmalen $a_1, a_2, \dots, a_{\frac{m}{2}}$, falls m gerade ist bzw. in den ersten $\frac{m-1}{2}$ Merkmalen $a_1, a_2, \dots, a_{\frac{m-1}{2}}$, falls m ungerade ist, erzeugt. Jede Dummy-Variable wird dabei als eigenes Merkmal aufgefasst, sodass $m = k + 1$ gilt. Des Weiteren wird die Ausfallquote q, d. h. der Anteil der fehlenden Werte in den Merkmalen, die vom Ausfall betroffen sind, von 5 % bis 25 % in 5 % Schritten variiert.

Beim MCAR Ausfallmechanismus werden in jedem vom Ausfall betroffenen Merkmal zufällig Werte gelöscht. Die Anzahl der Werte, die in den einzelnen Merkmalen zu löschen ist, entspricht dabei dem Produkt aus der Objektanzahl n und der Ausfallquote q. Falls die Anzahl der zu löschenden Werte nicht ganzzahlig ist, erfolgt eine Aufrundung auf die nächst größere ganze Zahl. So werden z. B. beim Autodatensatz und der Ausfallquote $q = 5\,\%$ in jedem vom Ausfall betroffenen Merkmal 20 Werte gelöscht, da das Produkt aus Ausfallquote und Objektanzahl den Wert $392 * 0{,}05 = 19{,}6$ annimmt.[8]

Für den MAR Ausfallmechanismus wird für jedes Merkmal mit fehlenden Werten ein weiteres Merkmal bestimmt, das den Ausfall in diesem Merkmal steuert. Die Zuordnung der Merkmale erfolgt, indem das letzte Merkmal den Ausfall im ersten Merkmal, das vorletzte Merkmal den Ausfall im zweiten Merkmal usw. steuert. Abbildung 1 zeigt eine schematische Darstellung dieser Zuordnung für eine ungerade Merkmalsanzahl m. Die schraffierten Flächen stellen dabei die Merkmale mit fehlenden Werten dar. Bei einer ungeraden Merkmalsanzahl ist das Merkmal $a_{\frac{m+1}{2}}$ somit weder vom Ausfall betroffen, noch steuert dieses den Ausfall in einem anderen Merkmal.

[8] Die Auswirkungen der durch dieses Vorgehen resultierenden (geringfügigen) Erhöhung der tatsächlich vorliegenden Ausfallquote sollen im Rahmen dieser Arbeit vernachlässigt werden.

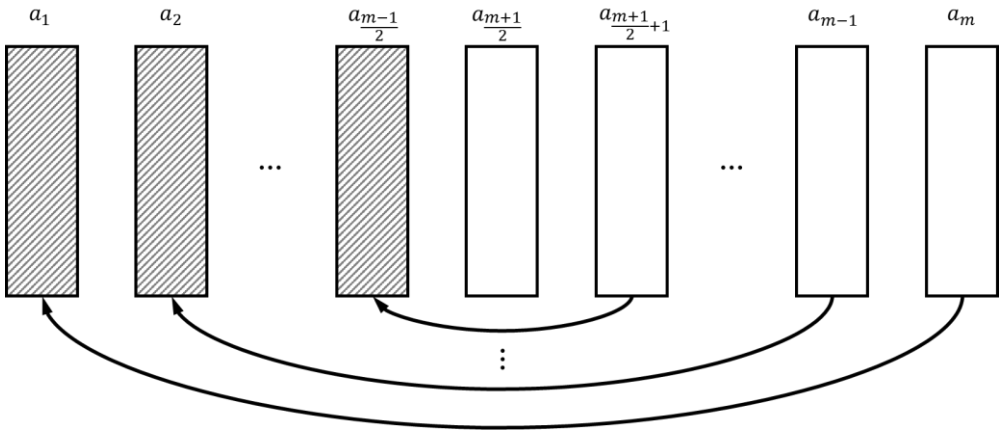

Abbildung 1: *Schematische Darstellung des MAR Ausfallmechanismus mit ungerader Merkmalsanzahl (Quelle: Eigene Darstellung in Anlehnung an Rockel (2018, S. 3))*

Unter Verwendung des MAR Ausfallmechanismus erfolgt die Steuerung des Ausfalls im Merkmal j ausgehend von dem Median $a_{\tilde{\jmath},med}$ des ausfallsteuernden Merkmals $\tilde{\jmath}$. Dabei stellt $N_{\tilde{\jmath},\leq med}$ die Indexmenge aller Objekte, deren Wert im Merkmal $\tilde{\jmath}$ kleiner-gleich dem Median $a_{\tilde{\jmath},med}$ ist, dar (vgl. Gl. (4)). Die Indexmenge aller Objekte, deren Wert im Merkmal $\tilde{\jmath}$ größer als der Median $a_{\tilde{\jmath},med}$ ist, wird als $N_{\tilde{\jmath},>med}$ bezeichnet (vgl. Gl. (5)). Der MAR Ausfallmechanismus wählt $\left|N_{\tilde{\jmath},\leq med}\right| * \tilde{q}$ Indizes aus $N_{\tilde{\jmath},\leq med}$ und $3 * \left|N_{\tilde{\jmath},>med}\right| * \tilde{q}$ Indizes aus $N_{\tilde{\jmath},>med}$ aus und löscht für die ausgewählten Indizes die Werte im Merkmal j. Dabei kann der gemäß Gl. (6) berechnete Parameter $\tilde{q}$ als die Ausfallwahrscheinlichkeit eines Objektes der Indexmenge $N_{\tilde{\jmath},\leq med}$ im Merkmal j interpretiert werden. Dieses Vorgehen ermöglicht es, die gewünschte Ausfallquote q zu erreichen, auch wenn eine Einteilung in zwei gleich große Gruppen anhand des Medians des ausfallsteuernden Merkmals $\tilde{\jmath}$ nicht möglich ist. Falls die Anzahl der zu löschenden Werte nicht ganzzahlig ist, erfolgt analog zum MCAR Ausfallmechanismus eine Aufrundung auf die nächst größere ganze Zahl.[9]

$$N_{\tilde{\jmath},\leq med} = \left\{i : a_{i\tilde{\jmath}} \leq a_{\tilde{\jmath},med}\right\} \tag{4}$$

$$N_{\tilde{\jmath},>med} = \left\{i : a_{i\tilde{\jmath}} > a_{\tilde{\jmath},med}\right\} \tag{5}$$

$$\tilde{q} = \frac{q * n}{\left|N_{\tilde{\jmath},\leq med}\right| + 3 * \left|N_{\tilde{\jmath},>med}\right|} \tag{6}$$

[9] Analog zum MCAR Ausfallmechanismus sollen die Auswirkungen der durch dieses Vorgehen resultierenden (geringfügigen) Erhöhung der tatsächlich vorliegenden Ausfallquote beim MAR und dem nachfolgenden MNAR Ausfallmechanismus im Rahmen dieser Arbeit vernachlässigt werden. Des Weiteren kann infolge der Aufrundung auch eine (geringfügige) Abweichung des Verhältnisses der zu löschenden Werte, die kleiner-gleich bzw. größer als der Median sind, resultieren.

In Anlehnung an Rockel (2018, S. 4) kann dieser Ausfallmechanismus auch als MAR1:3 Ausfallmechanismus bezeichnet werden, da das Verhältnis zwischen der Ausfallwahrscheinlichkeit der Werte, deren Ausprägung im Merkmal $\tilde{\jmath}$ kleiner-gleich dem Median ist, zu der Ausfallwahrscheinlichkeit der Werte, deren Ausprägung im Merkmal $\tilde{\jmath}$ größer als der Median ist, $1{:}\,3$ beträgt (vgl. Rockel 2018, S. 3f.).[10]

Der MNAR Ausfallmechanismus ähnelt in seiner Struktur dem MAR Ausfallmechanismus. Im Unterschied zum MAR Ausfallmechanismus wird der Ausfall im Merkmal j allerdings vom Merkmal j selbst gesteuert. Dabei stellen $N_{j,\leq med}$ bzw. $N_{j,>med}$ die Indexmenge aller Objekte, deren Wert im Merkmal j kleiner-gleich bzw. größer als der Median $a_{j,med}$ ist, dar (vgl. Gl. (7) und (8)). Der MNAR Ausfallmechanismus wählt $\left|N_{j,\leq med}\right| * \tilde{q}$ Indizes aus $N_{j,\leq med}$ und $3 * \left|N_{j,>med}\right| * \tilde{q}$ Indizes aus $N_{j,>med}$ aus und löscht die Werte im Merkmal j für die ausgewählten Indizes. Dieser Ausfallmechanismus kann analog zum MAR1:3 Ausfallmechanismus auch als MNAR1:3 Ausfallmechanismus bezeichnet werden. Falls die Anzahl der zu löschenden Werte nicht ganzzahlig ist, erfolgt erneut eine Aufrundung.

$$N_{j,\leq med} = \left\{i{:}\, a_{ij} \leq a_{j,med}\right\} \tag{7}$$

$$N_{j,>med} = \left\{i{:}\, a_{ij} > a_{j,med}\right\} \tag{8}$$

$$\tilde{q} = \frac{q * n}{\left|N_{j,\leq med}\right| + 3 * \left|N_{j,>med}\right|} \tag{9}$$

Die Behandlung der fehlenden Daten wird anhand der folgenden, in Kapitel 2.2 vorgestellten, MD-Verfahren untersucht:

- Complete-case analysis (CCA)
- Imputation durch den arithmetischen Mittelwert (IMW)
- Deterministische Regressionsimputation (DRI)
- Stochastische Regressionsimputation (SRI)
- Simple Random Hot-Deck Imputation (SRHDI)
- Nearest Neighbor Hot-Deck Imputation (NNHDI)

[10] Zur Verwendung eines MAR1:V Ausfallmechanismus, bei dem das Verhältnis zwischen den Ausfallwahrscheinlichkeiten den Wert V anstatt 3 annimmt, sind die folgenden Anpassungen erforderlich: Der MAR1:V Ausfallmechanismus wählt $V * \left|N_{j,>med}\right| * \tilde{q}$ Indizes aus $N_{j,>med}$ aus und der Parameter $\tilde{q}$ berechnet sich gemäß: $\tilde{q} = \frac{q*n}{\left|N_{j,\leq med}\right| + V * \left|N_{j,>med}\right|}$.

Für die Umsetzung der CCA wird das R-Paket stats (Version 3.6.3) verwendet. Die Funktionen für die IMW und SRHDI werden in R implementiert. Die Anwendung der DRI und SRI erfolgt mittels des R-Pakets mice (Version 3.8.0) (van Buuren und Groothuis-Oudshoorn 2011). Zur Umsetzung der NNHDI wird das R-Paket HotDeckImputation (Joenssen 2015a) in der Version 1.1.0 verwendet.

Die Bewertung der Auswirkungen von fehlenden Daten und der Vergleich der MD-Verfahren erfolgen anhand der drei Bewertungskriterien: relative Abweichung der Regressionskoeffizienten, Abweichung des korrigierten Bestimmtheitsmaßes und Abweichung der prognostizierten Werte der abhängigen Variable. Dabei wird für die Bewertung der Abweichung der Regressionskoeffizienten der Betrag der relativen Abweichung zwischen den Regressionskoeffizienten des Regressionsmodells mit vollständigen Daten $\hat{\beta}_{j,vollständig}$ und denen des Regressionsmodells mit fehlenden Daten $\hat{\beta}_{j,MD}$ verwendet (vgl. Gl. (10)). Der Wert aus Gl. (10) kann als die durchschnittliche relative Abweichung der Regressionskoeffizienten im Vergleich zum Regressionsmodell ohne fehlende Daten interpretiert werden.

$$\frac{1}{k+1}\sum_{j=0}^{k}\left|\frac{\hat{\beta}_{j,MD}-\hat{\beta}_{j,vollständig}}{\hat{\beta}_{j,vollständig}}\right| \tag{10}$$

Die Abweichung des korrigierten Bestimmtheitsmaßes wird gemäß Gl. (11) bestimmt. Da nur eine einzelne Differenz bestimmt wird, entspricht der Betrag der absoluten Abweichung dem Root Mean Squared Error (RMSE) zwischen dem korrigierten Bestimmtheitsmaß des Regressionsmodells mit vollständigen Daten und dem Regressionsmodell mit fehlenden Daten. Dabei können die korrigierten Bestimmtheitsmaße der Regressionsmodelle mit vollständigen Daten $\bar{R}^2_{vollständig}$ der Tabelle 1 entnommen werden.

$$\sqrt{\left(\bar{R}^2_{MD}-\bar{R}^2_{vollständig}\right)^2}=\left|\bar{R}^2_{MD}-\bar{R}^2_{vollständig}\right| \tag{11}$$

Zur Beurteilung der prognostizierten Werte wird der RMSE zwischen den wahren Werten y_i und den, nach Anwendung der MD-Verfahren mit den multiplen linearen Regressionsmodell prognostizierten, Werten $\hat{y}_{i,MD}$ bestimmt. Die Berechnung der prognostizierten Werte $\hat{y}_{i,MD}$ erfolgt ausgehend von den geschätzten Regressionskoeffizienten $\hat{\beta}_{j,MD}$ und den wahren

Werten der unabhängigen Variablen $\boldsymbol{X}$ gemäß Gl. (13). Zur Verbesserung der Vergleichbarkeit der Datensätze wird die Differenz der prognostizierten und beobachteten Werte durch die Stichprobenstandardabweichung s_y der abhängigen Variable y geteilt (vgl. Gl. (12)).

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\hat{y}_{i,MD}-y_i}{s_y}\right)^2} \tag{12}$$

$$\hat{\boldsymbol{y}}_{\boldsymbol{MD}} = \boldsymbol{X}\,\hat{\boldsymbol{\beta}}_{\boldsymbol{MD}} \tag{13}$$

Dieser Simulationsstudie liegt ein vollständiger Versuchsplan zugrunde, sodass alle Faktorstufenkombinationen der Einflussgrößen simuliert werden. Um eine Verfälschung der Untersuchungsergebnisse zu verhindern, werden alle Faktorstufenkombinationen 1000 mal simuliert (vgl. Rockel 2017, S. 12f.). Die im weiteren Verlauf dargestellten Ergebnisse sind jeweils Mittelwerte der $I = 1000$ Wiederholungen. Das resultierende Studiendesign ist in Abbildung 2 dargestellt. Aus dieser Übersicht geht hervor, dass insgesamt 90 verschiedene Kombinationen aus Datensätzen und fehlenden Daten im Rahmen dieser Simulationsstudie betrachtet werden.[11]

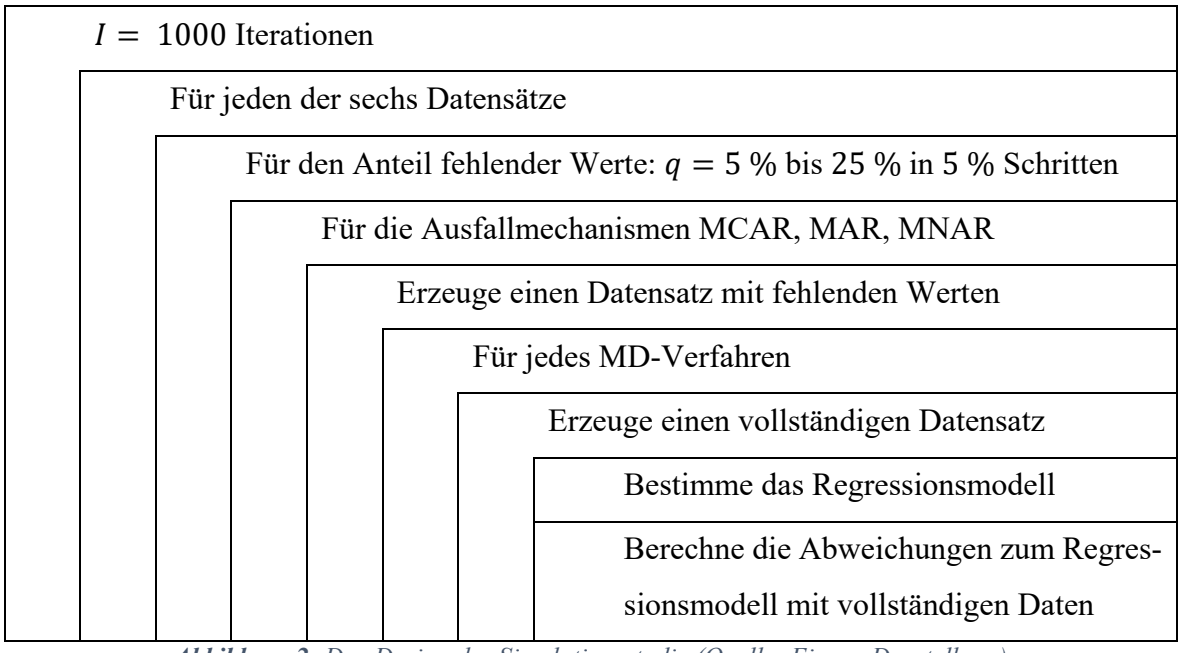

Abbildung 2: *Das Design der Simulationsstudie (Quelle: Eigene Darstellung)*

[11] Ausgehend von sechs verschiedenen Datensätzen, fünf verschiedenen Ausfallquoten und drei verschiedenen Ausfallmechanismen ergeben sich $6 * 5 * 3 = 90$ verschiedene Kombinationen.

4 Ergebnisse der Simulationsstudie

Die grafische Darstellung der Simulationsergebnisse erfolgt in Anlehnung an Rockel (2018, S. 7ff.). Zur besseren visuellen Unterscheidung der MD-Verfahren wird ein eingeschränkter Wertebereich der Ordinatenachse verwendet. Des Weiteren wird auf eine Durchführung der CCA im Rahmen des Hittersdatensatzes verzichtet, da infolge der erheblichen Reduzierung der Objektanzahl eine eindeutige Schätzung der Regressionskoeffizienten nicht in jedem Fall möglich ist. Bei einer Ausfallquote von $q = 25\ \%$ reduziert sich die erwartete Anzahl der vollständigen Objekte auf $(1 - 0{,}25)^{10} = 5{,}63\ \%$ des ursprünglichen Niveaus, wenn die fehlenden Werte zufällig und unabhängig voneinander in den vom Ausfall betroffenen Merkmalen auftreten. Somit würden für den Hittersdatensatz bei einer Ausfallquote von $q = 25\ \%$ im Mittel lediglich $(1 - 0{,}25)^{10} * 263 = 14{,}81$ vollständige Objekte zur Schätzung der 21 Regressionskoeffizienten zur Verfügung stehen.

4.1 Auswirkungen auf die Regressionskoeffizienten

Die Resultate der Simulationsstudie zur Beurteilung der relativen Abweichung der Regressionskoeffizienten sind in Abbildung 3 dargestellt. Diese Abbildung ist zeilenweise in die drei in Kapitel 3 beschriebenen Ausfallmechanismen und spaltenweise in die sechs verwendeten Datensätze unterteilt. Für jede Kombination der drei Ausfallmechanismen und sechs Datensätze wird auf der Abszisse die Ausfallquote q und auf der Ordinate die durchschnittliche relative Abweichung der Regressionskoeffizienten dargestellt. Ein höherer Wert auf der Ordinate entspricht dabei einer größeren relativen Abweichung der Regressionskoeffizienten. Das „beste“ Ergebnis erzielt demzufolge das MD-Verfahren, welches bei einer gegebenen Faktorstufenkombination die geringste durchschnittliche relative Abweichung der Regressionskoeffizienten erreicht. Eine Faktorstufenkombination entspricht in diesem Zusammenhang der Kombination aus einem Datensatz, einem Ausfallmechanismus und einer Ausfallquote. Folglich erreicht beispielsweise die NNHDI bei dem Autodatensatz unter einem MCAR Ausfallmechanismus und der Ausfallquote $q = 25\ \%$ mit einer durchschnittlichen relativen Abweichung der Regressionskoeffizienten von 32,78 % ein besseres Ergebnis als die IMW mit 65,53 %.

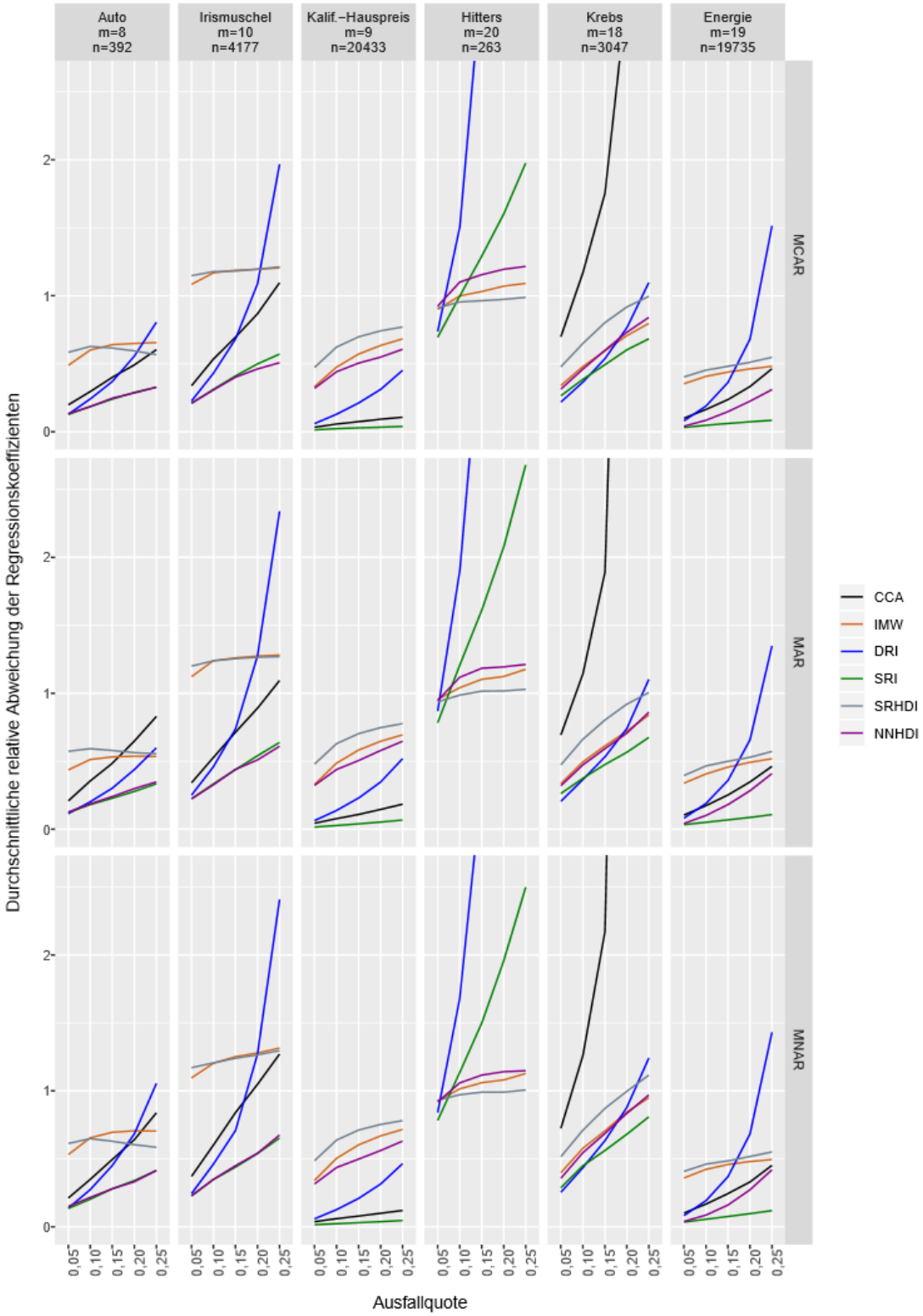

Abbildung 3: *Durchschnittliche relative Abweichung der Regressionskoeffizienten (Quelle: Eigene Darstellung)*

Bei dem Auto- und Irismuscheldatensatz (dargestellt im linken Teil der Abbildung 3) führen die SRI und NNHDI zu den besten Ergebnissen. Mit Ausnahme des Hittersdatensatzes (geringste Objekt- und höchste Merkmalsanzahl) erzielt die SRI auch bei den anderen Datensätzen überwiegend die besten Ergebnisse, während die NNHDI mehrheitlich Ergebnisse im Mittelfeld erreicht. Die DRI erzielt bei sehr geringen Ausfallquoten ($q \leq 10\,\%$) mehrfach vergleichbare, vereinzelt sogar geringfügig bessere, Ergebnisse als die SRI, verschlechtert sich aber über alle Datensätze deutlich mit zunehmender Ausfallquote, sodass sie bei einer Ausfallquote von $q = 25\,\%$ überwiegend zu den schlechtesten Verfahren zählt. Bei dem Hittersdatensatz tritt diese ausgeprägte Verschlechterung mit zunehmender Ausfallquote auch bei der SRI auf, sodass die SRI für Ausfallquoten $q > 10\,\%$ schlechtere Ergebnisse erzielt als die IMW, SRHDI und NNHDI. Die besseren Ergebnisse der SRI gegenüber der DRI spiegeln die Überlegung wider, dass die SRI besser geeignet ist, um Korrelationen im Datensatz zu erhalten.

Während die CCA bei den Datensätzen mit sehr vielen Objekten (Kalifornien-Hauspreisdatensatz und Energiedatensatz) gute Ergebnisse erzielt, erreicht sie bei den anderen Datensätzen in Abhängigkeit von der Ausfallquote q nur mäßige bis schlechte Ergebnisse. Dabei kann insbesondere bei dem Krebsdatensatz eine deutliche Verschlechterung mit zunehmender Ausfallquote beobachtet werden.

Die IMW und SRHDI zählen außer bei dem Hittersdatensatz überwiegend zu den schlechtesten Verfahren, weisen aber im Vergleich zu den anderen untersuchten MD-Verfahren mehrheitlich nur eine geringfügige Verschlechterung mit zunehmender Ausfallquote auf. Des Weiteren erreichen diese zwei Verfahren bei dem Hittersdatensatz für Ausfallquoten oberhalb von $q = 10\,\%$ die besten Ergebnisse, gefolgt von der NNHDI. Die IMW erzielt zudem beim Krebsdatensatz Ergebnisse im Mittelfeld, vergleichbar mit der NNHDI.

Allgemein zeigt sich, dass alle Verfahren mit zunehmender Ausfallquote zu größeren relativen Abweichungen der geschätzten Regressionskoeffizienten führen. Eine Ausnahme besteht lediglich bei Anwendung der SRHDI zur Behandlung der fehlenden Werte des Autodatensatzes. Bezüglich des Ausfallmechanismus werden unter einem MAR und MNAR Ausfallmechanismus vermehrt geringfügig schlechtere Ergebnisse erzielt als unter einem MCAR Ausfallmechanismus. Des Weiteren scheint die Mehrheit der Verfahren bei Datensätzen mit sehr vielen Objekten (Kalifornien-Hauspreisdatensatz und Energiedatensatz) zu

ähnlicheren Schätzwerten der Regressionskoeffizienten zu führen. Der Einfluss der Merkmalsanzahl ist hingegen nicht eindeutig. Mit Ausnahme des Hittersdatensatzes zeigt sich, dass die SRI, insbesondere bei höheren Ausfallquoten, mehrheitlich zu den ähnlichsten Schätzungen der Regressionskoeffizienten führt.

4.2 Auswirkungen auf das korrigierte Bestimmtheitsmaß

Abbildung 4 zeigt die Resultate der Simulationsstudie zur Beurteilung der Abweichung des korrigierten Bestimmtheitsmaßes $\bar{R}^2$. Diese Abbildung ist analog zu Abbildung 3 aufgebaut. Der einzige Unterschied zwischen diesen beiden Abbildungen besteht darin, dass nun der RMSE zwischen dem korrigierten Bestimmtheitsmaß des Regressionsmodells mit vollständigen Daten und dem Regressionsmodell mit fehlenden Daten auf der Ordinate abgetragen ist. Analog zur Abweichung der Regressionskoeffizienten erzielt das MD-Verfahren das „beste“ Ergebnis, welches bei einer gegebenen Faktorstufenkombination den geringsten RMSE erreicht. Am Beispiel des Autodatensatzes unter einem MCAR Ausfallmechanismus und der Ausfallquote $q = 25\ \%$ erreicht die NNHDI mit einem RMSE von $0{,}0090$ erneut ein besseres Ergebnis als die IMW mit einem RMSE von $0{,}0592$.

Mit Ausnahme des Hittersdatensatzes führt die SRI bei allen Datensätzen und Ausfallmechanismen wiederholt zu den besten Ergebnissen. Bei dem Hittersdatensatz tritt erneut die bereits bei der Abweichung der Regressionskoeffizienten beobachtete, deutliche Verschlechterung mit zunehmender Ausfallquote auf. Dies führt dazu, dass die NNHDI beim Hittersdatensatz für eine Ausfallquote $q > 15\ \%$ (bei einem MAR Ausfallmechanismus für $q > 10\ \%$) bessere Ergebnisse als die SRI erzielt. Auch die IMW erreicht bei diesem Datensatz für eine Ausfallquote $q > 20\ \%$ (bei einem MAR Ausfallmechanismus für $q > 15\ \%$) bessere Ergebnisse als die SRI. Die DRI führt überwiegend zu Ergebnissen im Mittelfeld und zeigt insbesondere bei dem Hitters- und Energiedatensatz eine deutliche Verschlechterung mit zunehmender Ausfallquote, sodass die DRI für eine Ausfallquote $q > 15\ \%$ zu den schlechtesten Ergebnissen beim Hittersdatensatz führt.

Während die NNHDI überwiegend zu den besseren Verfahren zählt und lediglich beim Kalifornien-Hauspreisdatensatz ein merklich schlechteres Ergebnis als die DRI und CCA erzielt, stellt die SRHDI außer bei dem Hittersdatensatz stets das schlechteste Verfahren dar.

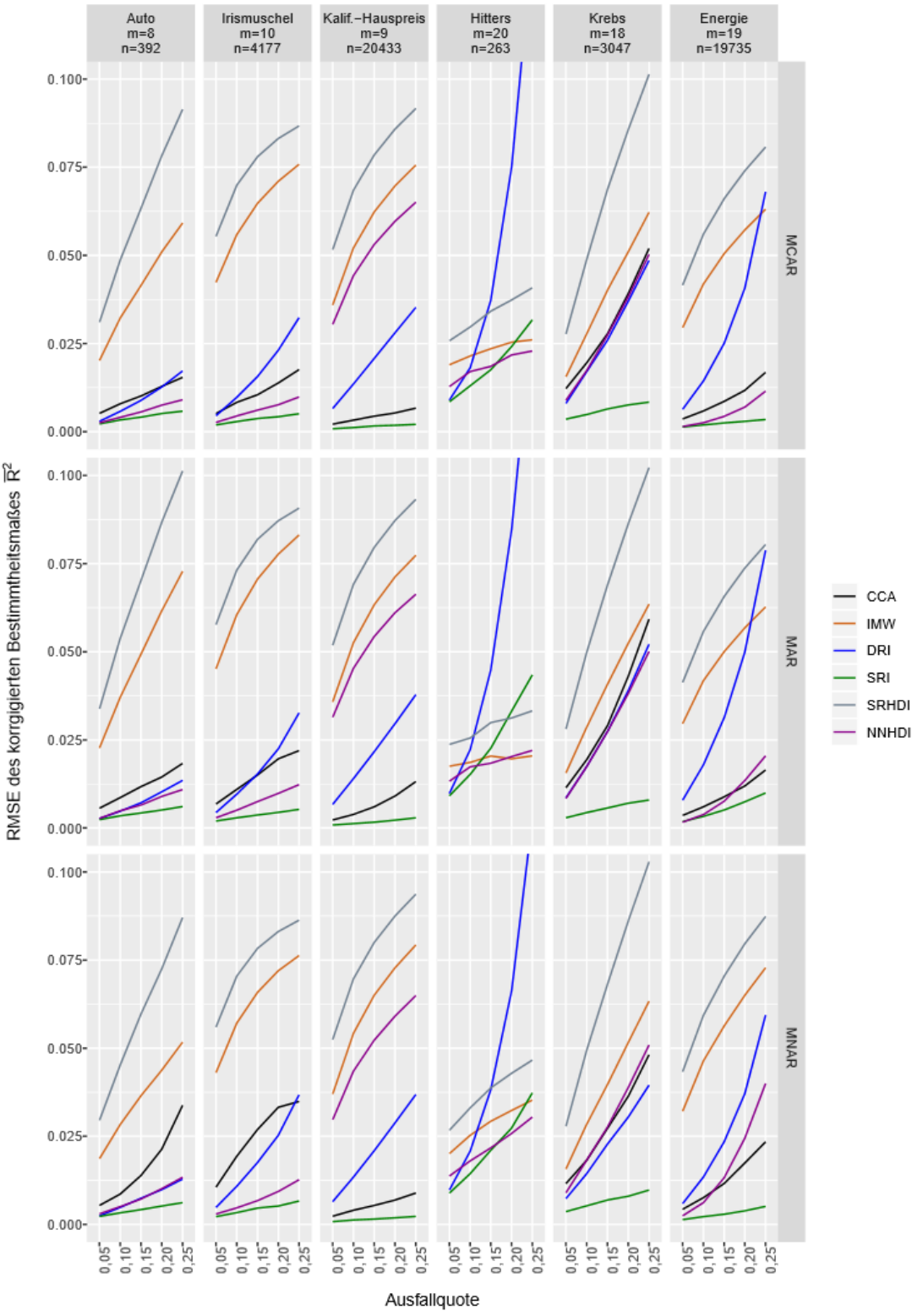

Abbildung 4: *RMSE der Abweichung des korrigierten Bestimmtheitsmaßes (Quelle: Eigene Darstellung)*

Beim Hittersdatensatz zählt die SRHDI ebenfalls zu den schlechteren Verfahren. Die (simultane) NNHDI führt somit zu ähnlicheren Schätzungen des korrigierten Bestimmtheitsmaßes als die (sequenzielle) SRHDI. Auch die IMW erzielt abgesehen vom Hittersdatensatz nahezu durchgehend die schlechtesten Ergebnisse nach der SRHDI, erreicht aber beim Hittersdatensatz bei geringen Ausfallquoten Ergebnisse im Mittelfeld und bei höheren Ausfallquoten gute Ergebnisse. Die CCA erzielt bei den beiden Datensätzen mit sehr vielen Objekten erneut gute Ergebnisse und erreicht bei den anderen Datensätzen überwiegend Ergebnisse im Mittelfeld.

Insgesamt zeigt sich, dass die Abweichung des korrigierten Bestimmtheitsmaßes bei allen Verfahren mit steigender Ausfallquote zunimmt. Die einzige Ausnahme scheinen hierbei die Ergebnisse der IMW zur Behandlung der fehlenden Werte des Hittersdatensatzes unter einem MAR Ausfallmechanismus zu sein, die sich teilweise mit zunehmendem Anteil fehlender Werte leicht verbessern. Des Weiteren scheint die Mehrheit der Verfahren unter einem MCAR Ausfallmechanismus überwiegend geringfügig bessere Ergebnisse zu erzielen als unter einem MAR oder MNAR Ausfallmechanismus. Eine überwiegende Verbesserung der Ergebnisse mit zunehmender Objektanzahl kann nur bei der CCA und SRI beobachtet werden. Der Einfluss der Merkmalsanzahl ist erneut nicht eindeutig. So scheint z. B. die NNHDI in Relation zu den Ergebnissen der anderen MD-Verfahren überwiegend von einer höheren Merkmalsanzahl zu profitieren, während sich beispielsweise die DRI infolgedessen mehrheitlich verschlechtert. Mit Ausnahme des Hittersdatensatzes zeigt sich über alle anderen Datensätze hinweg, dass die SRI, insbesondere bei höheren Ausfallquoten, auch bei dem korrigierten Bestimmtheitsmaß zu den geringsten Abweichungen führt.

4.3 Auswirkungen auf die prognostizierten Werte

Die Resultate der Simulationsstudie zur Beurteilung der Abweichung der prognostizierten Werte sind in Abbildung 5 dargestellt. Im Unterschied zu den vorangegangenen Abbildungen ist nun der RMSE zwischen den beobachteten und prognostizierten Werten der abhängigen Variable auf der Ordinate abgetragen. Analog zu den vorangegangenen Untersuchungen erreicht das MD-Verfahren das „beste“ Ergebnis, welches bei einer gegebenen Faktorstufenkombination den geringsten RMSE erreicht. Somit erzielt die NNHDI am Beispiel des Autodatensatzes unter einem MCAR Ausfallmechanismus und der Ausfallquote $q = 25\,\%$

auch bei diesem Bewertungskriterium mit einem RMSE von 0,4215 ein besseres Ergebnis als die IMW mit einem RMSE von 0,4641.

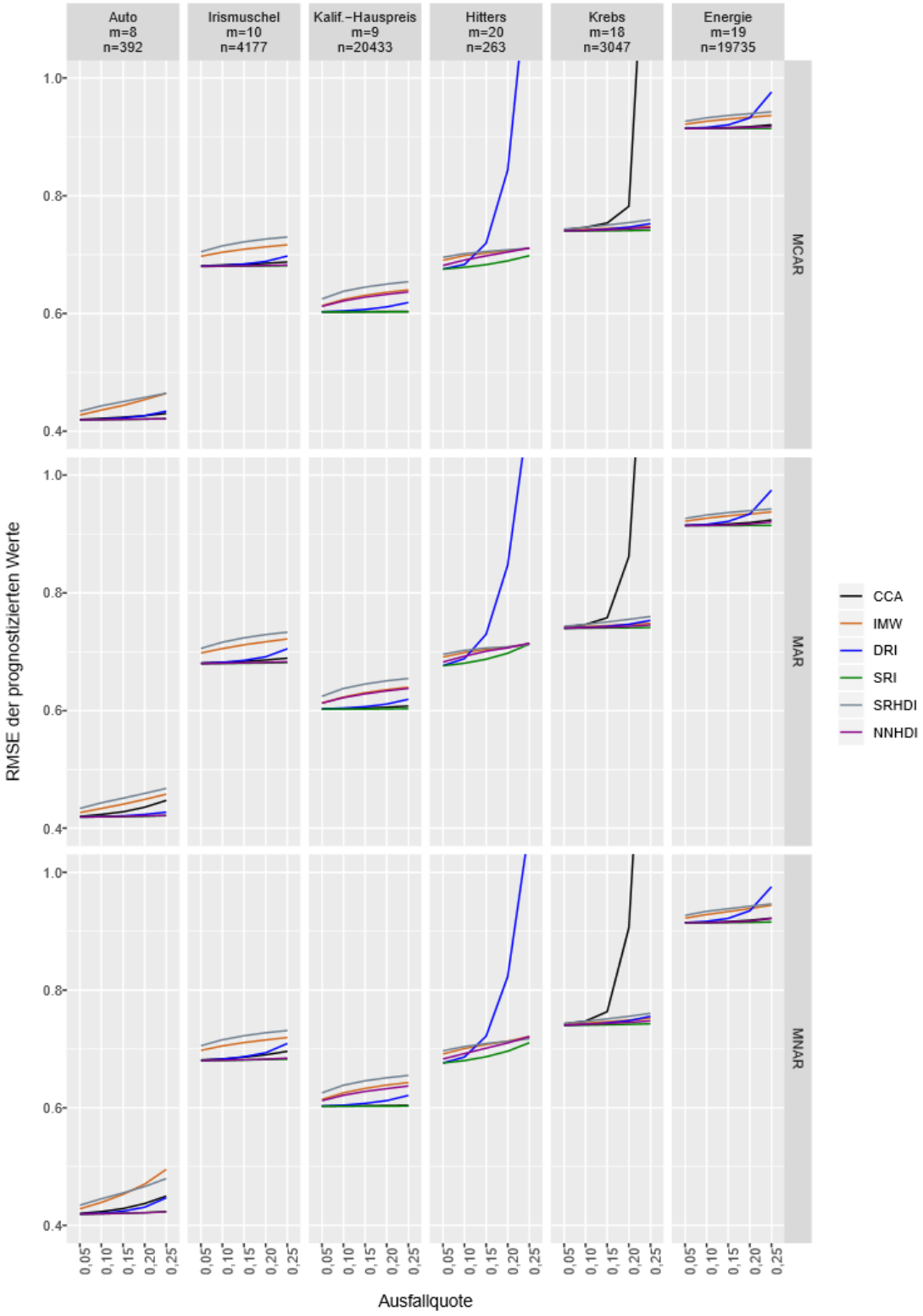

Abbildung 5: *RMSE zwischen den beobachteten und prognostizierten Werten der abhängigen Variable (Quelle: Eigene Darstellung)*

Ergänzend zu Abbildung 5 zeigen die Abbildungen 6 – 8 eine auf die jeweiligen Datensätze zugeschnittene Darstellung der Simulationsergebnisse zur besseren visuellen Differenzierung der MD-Verfahren. Diese unterscheiden sich dahingehend von Abbildung 5, dass ein kleinerer Wertebereich auf der Ordinate dargestellt wird.

Im Gegensatz zur Untersuchung der Abweichungen der Regressionskoeffizienten und des korrigierten Bestimmtheitsmaßes führt die SRI nun über alle Datensätze und Ausfallmechanismen hinweg, d. h. auch bei dem Hittersdatensatz, nahezu durchgehend zu den besten Ergebnissen. Die NNHDI erzielt bei dem Auto- und Irismuscheldatensatz (dargestellt im linken Teil der Abbildung 5) nahezu identische Ergebnisse wie die SRI und erreicht auch bei den anderen Datensätzen erneut gute Ergebnisse. Lediglich beim Kalifornien-Hauspreisdatensatz erreicht die NNHDI ein deutlich schlechteres Ergebnis als die DRI und CCA. Während die DRI bei niedrigen Ausfallquoten ($q \leq 10\,\%$) oft gute Ergebnisse erreicht, verschlechtert sie sich zunehmend mit steigender Ausfallquote, sodass sie bei höheren Ausfallquoten ($q \geq 15\,\%$) überwiegend Ergebnisse im Mittelfeld erreicht. Dabei tritt insbesondere bei dem Hitters- und Energiedatensatz eine deutliche Verschlechterung auf. Dies führt dazu, dass die DRI bei dem Hittersdatensatz für Ausfallquoten $q \geq 15\,\%$ und bei dem Energiedatensatz für Ausfallquoten $q > 20\,\%$ die schlechtesten Ergebnisse erzielt. Die SRI weist somit erneut Vorteile gegenüber der DRI auf.

Die CCA erzielt bei den beiden Datensätzen mit sehr vielen Objekten erneut gute Ergebnisse und erreicht bei dem Kalifornien-Hauspreisdatensatz unter den MCAR und MNAR Ausfallmechanismen nahezu identische Ergebnisse wie die SRI. Allerdings kann bei dem Krebsdatensatz erneut eine deutliche Verschlechterung mit zunehmender Ausfallquote beobachtet werden, sodass die CCA bei diesem Datensatz für Ausfallquoten $q > 10\,\%$ zu den schlechtesten Ergebnissen führt. Bei dem Auto- und Irismuscheldatensatz erreicht die CCA wiederholt überwiegend Ergebnisse im Mittelfeld.

Die SRHDI zählt über alle Datensätze und Ausfallmechanismen hinweg überwiegend zu den schlechtesten Verfahren. Die einzige Ausnahme scheint hierbei die Behandlung der fehlenden Werte des Hittersdatensatzes bei höheren Ausfallquoten $q \geq 20\,\%$ zu sein, bei denen die SRHDI vergleichbare Ergebnisse wie die SRI und NNHDI erzielt. Die (simultane) NNHDI erweist sich somit auch bei den prognostizierten Werten der abhängigen Variable als vorteilhafter als die (sequenzielle) SRHDI. Die IMW führt erneut überwiegend zu

schlechten Ergebnissen. Lediglich beim Krebsdatensatz erreicht sie Ergebnisse im Mittelfeld.

Eine mögliche Ursache für die vergleichsweise niedrigen bzw. hohen Abweichungen der prognostizierten Werte beim Auto- bzw. Energiedatensatz ist anhand der Bestimmtheitsmaße der Regressionsmodelle mit vollständigen Daten ersichtlich. Während beim Autodatensatz ca. 82,41 % der Varianz der abhängigen Variable erklärt werden, können beim Energiedatensatz lediglich ca. 16,44 % erklärt werden (vgl. Tabelle 1). Infolgedessen weisen auch die, ausgehend von den ursprünglich unvollständigen Daten gebildeten, Regressionsmodelle vergleichsweise niedrigere bzw. höhere Abweichungen der prognostizierten Werte auf als die Regressionsmodelle der anderen vier Datensätzen.[12]

Erneut zeigt sich, dass die Abweichung der prognostizierten Werte bei allen Verfahren mit steigender Ausfallquote zunimmt, die Verschlechterung aber zwischen den Verfahren und Datensätzen stark variiert. Dabei erweist sich insbesondere die SRI mit Ausnahme des Hittersdatensatzes als robust gegenüber einer Erhöhung der Ausfallquote. Der Einfluss der Objektanzahl und des Ausfallmechanismus ist hingegen nicht eindeutig. Die Mehrheit der Verfahren scheint aber unter einem MCAR Ausfallmechanismus erneut geringfügig bessere Ergebnisse zu erreichen als unter einem MAR oder MNAR Ausfallmechanismus. Des Weiteren scheint eine höhere Merkmalsanzahl bei einer ähnlichen Objektanzahl tendenziell zu einer größeren Abweichung der prognostizierten Werte zu führen. Insbesondere die CCA weist Nachteile infolge einer höheren Merkmalsanzahl und der damit verbundenen Reduzierung der Anzahl vollständiger Objekte auf. Darüber hinaus zeigt sich erneut, dass die SRI, insbesondere bei höheren Ausfallquoten, zu den geringsten Abweichungen der prognostizierten Werte führt.

[12] Die Bestimmtheitsmaße der Regressionsmodelle mit vollständigen Daten der anderen vier Datensätze liegen im Bereich von 45,28 % bis 63,69 % (vgl. Tabelle 1).

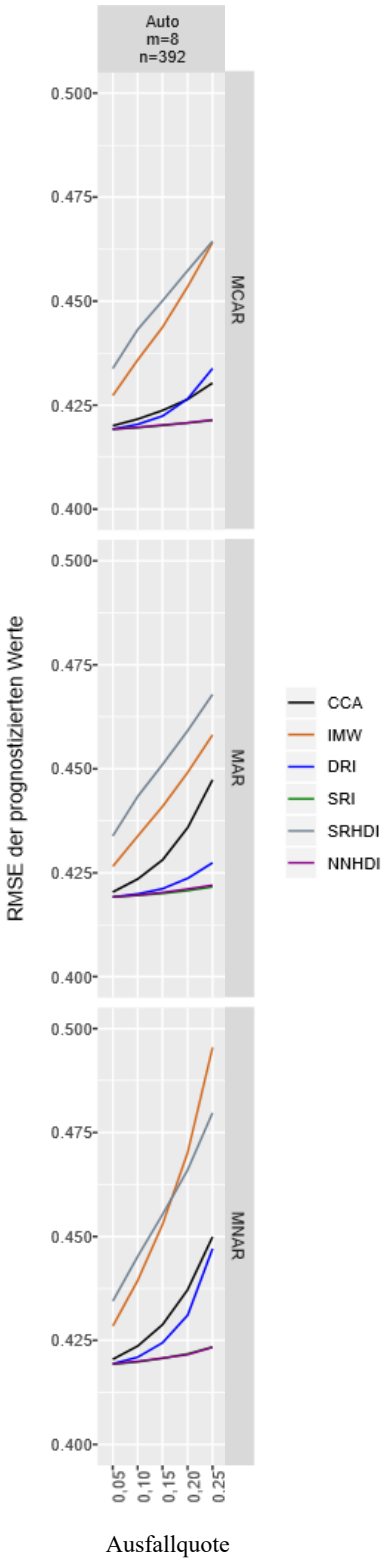

Abbildung 6: *RMSE zwischen den beobachteten und prognostizierten Werten der abhängigen Variable des Autodatensatzes (Quelle: Eigene Darstellung)*

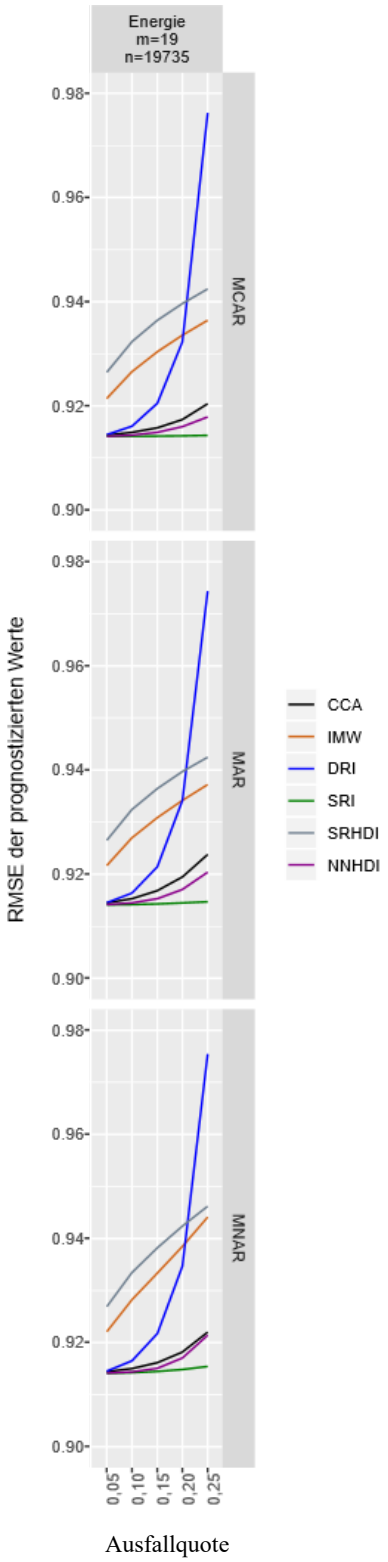

Abbildung 7: *RMSE zwischen den beobachteten und prognostizierten Werten der abhängigen Variable des Energiedatensatzes (Quelle: Eigene Darstellung)*

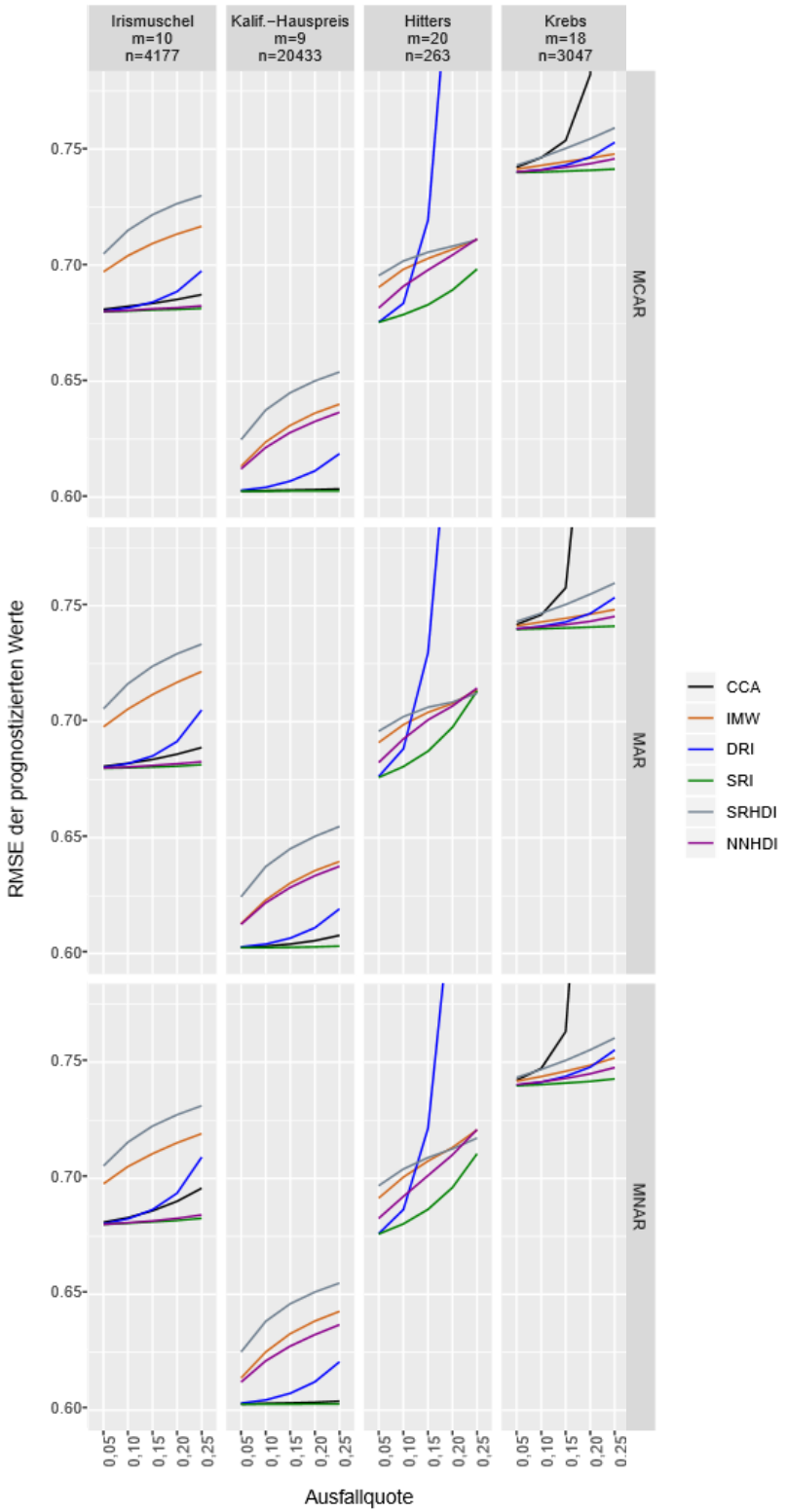

Abbildung 8: *RMSE zwischen den beobachteten und prognostizierten Werten der abhängigen Variable des Irismuschel-, Kalifornien-Hauspreis-, Hitters- und Krebsdatensatzes (Quelle: Eigene Darstellung)*

5 Auswertung und Interpretation der Ergebnisse

Auch wenn die SRI überwiegend die besten Ergebnisse erzielt, zeigt der Vergleich der Simulationsergebnisse über alle Bewertungskriterien und/oder Datensätze hinweg, dass es in der Simulationsstudie kein Verfahren gibt, das allen anderen Verfahren in jeder Hinsicht überlegen ist. Dabei sind die Auswirkungen der variierten Simulationsparameter auf die Bewertungskriterien in vielen Fällen ähnlich. Allerdings weisen die Variationen zum Teil auch erhebliche Unterschiede zwischen den einzelnen Bewertungskriterien und/oder Faktorstufen auf.

Bezüglich der fehlenden Daten zeigt sich, dass eine Erhöhung der **Ausfallquote** in nahezu allen Fällen zu einer Verschlechterung der Bewertungskriterien führt, die Verschlechterung der MD-Verfahren aber zum Teil erheblich zwischen den einzelnen Faktorstufenkombinationen und Bewertungskriterien variiert. Dabei weisen insbesondere die DRI bei dem Hittersund Energiedatensatz bei allen Bewertungskriterien und die CCA bei dem Krebsdatensatz bei zwei Bewertungskriterien deutliche Verschlechterungen mit zunehmender Ausfallquote auf. Die SRI zeigt sich in Relation zu den anderen untersuchten MD-Verfahren überwiegend robust gegenüber einer Erhöhung der Ausfallquote. Lediglich bei dem Hittersdatensatz weist die SRI eine deutliche Verschlechterung mit zunehmender Ausfallquote auf, die insbesondere bei der relativen Abweichung der Regressionskoeffizienten ersichtlich ist.

Des Weiteren ist z. B. bei der SRI und dem Hittersdatensatz ein deutlicher Unterschied zwischen den Bewertungskriterien zu erkennen. Während die SRI bezüglich der relativen Abweichung der Regressionskoeffizienten für Ausfallquoten $q > 10\,\%$ deutlich schlechtere Ergebnisse als die IMW, SRHDI und NNHDI erzielt, erreicht die SRI bezüglich der Abweichung der prognostizierten Werte der abhängigen Variable auch für hohe Ausfallquoten die besten Ergebnisse. Eine große Abweichung der Regressionskoeffizienten führt somit nicht zwangsläufig zu einer großen Abweichung der prognostizierten Werte der abhängigen Variable. Daraus folgt, dass das Ausmaß der Verschlechterung der MD-Verfahren infolge einer erhöhten Ausfallquote von den Eigenschaften des Datensatzes und dem Bewertungskriterium abhängt. Auch wenn die untersuchten Bewertungskriterien mehrfach ähnliche Verläufe aufweisen, ist es somit nicht möglich, mit Hilfe der Ergebnisse eines Bewertungskriteriums unmittelbar auf die Ergebnisse eines anderen Bewertungskriteriums zu schließen.

Der Einzeleffekt des **Ausfallmechanismus** ist weniger stark ausgeprägt als der Einzeleffekt der Ausfallquote. Tabelle 2 zeigt für jedes Bewertungskriterium einen Vergleich der Simulationsergebnisse unter Variation des Ausfallmechanismus. Dabei gibt der jeweilige Eintrag in Tabelle 2 für jedes Bewertungskriterium die Anzahl der 175 Vergleiche zweier Ausfallmechanismen an, bei denen unter dem erstgenannten Ausfallmechanismus bessere bzw. schlechtere Ergebnisse erzielt worden.[13] Dabei zeigt sich über alle Bewertungskriterien hinweg, dass in der Simulationsstudie unter einem MCAR Ausfallmechanismus überwiegend bessere Ergebnisse erreicht worden als unter einem MAR oder MNAR Ausfallmechanismus. Eine klare Entscheidung zugunsten des MAR Ausfallmechanismus gegenüber dem MNAR Ausfallmechanismus kann lediglich bei der Abweichung der prognostizierten Werte der abhängigen Variable getroffen werden. Da bei der Abweichung des korrigierten Bestimmtheitsmaßes und der Regressionskoeffizienten fast genauso häufig niedrigere Abweichungen unter einem MNAR Ausfallmechanismus wie unter einem MAR Ausfallmechanismus erreicht worden, erscheint eine klare Entscheidung zugunsten des MAR Ausfallmechanismus bei diesen Bewertungskriterien nur wenig zweckmäßig.

	MCAR vs. MAR		**MCAR vs. MNAR**		**MAR vs. MNAR**	
	MCAR besser	MCAR schlechter	MCAR besser	MCAR schlechter	MAR besser	MAR schlechter
Regressionskoeffizienten	126	49	158	17	104	71
Korrigiertes Bestimmtheitsmaß	131	44	129	46	90	85
Prognostizierte Werte	121	54	168	7	130	45

Tabelle 2: *Ausfallmechanismen im Vergleich (Quelle: Eigene Darstellung)*

Bei der Einflussgröße **Objektanzahl** ergibt sich kein eindeutiges Bild über alle MD-Verfahren und Bewertungskriterien, sodass eine differenzierte Betrachtung erforderlich ist. Während die Mehrheit der Verfahren bei der Schätzung der Regressionskoeffizienten von einer höheren Objektanzahl profitiert, kann ein solcher Effekt bei dem korrigierten Bestimmtheits-

[13] Für jedes Bewertungskriterium können 175 Vergleiche zwischen zwei verschiedenen Ausfallmechanismen durchgeführt werden. Die 175 Vergleiche resultieren aus den sechs verschiedenen Datensätzen, fünf verschiedenen Ausfallquoten und sechs verschiedenen MD-Verfahren unter Berücksichtigung, dass im Rahmen des Hittersdatensatzes auf eine Durchführung der CCA verzichtet wurde: $6 * 5 * 6 - 5 = 175$.

maß nur bei der CCA und SRI beobachtet werden. Bei der Prognose der Werte der abhängigen Variable scheint lediglich die CCA in Relation zu den anderen untersuchten MD-Verfahren von der Erhöhung der Objektanzahl zu profitieren, weist aber insbesondere beim Energiedatensatz eine absolute Verschlechterung auf. Bei den übrigen MD-Verfahren ist diese absolute Verschlechterung ebenfalls ersichtlich. Zudem scheint bei den übrigen MD-Verfahren eine einheitliche relative Auswirkung der erhöhten Objektanzahl nicht vorzuliegen. Mögliche Ursachen dafür können Wechselwirkungen zwischen einzelnen Faktoren und MD-Verfahren sowie weitere (nicht systematisch variierte) Eigenschaften der realen Datensätze, wie z. B. die Korrelation der Merkmale, sein.

Die (relative) Verbesserung der CCA infolge einer Erhöhung der Objektanzahl kann damit begründet werden, dass infolge des Ausschlusses aller unvollständigen Objekte ein großer Anteil der Objekte von der weiteren Untersuchung ausgeschlossen wird und somit nicht mehr zur Schätzung der Regressionskoeffizienten zur Verfügung steht. Die in der Literatur (vgl. Bankhofer und Vogel 2008, S. 228) geforderten Richtwerte bezüglich der Mindestanzahl an Objekten pro zu schätzenden Regressionskoeffizienten sind für Datensätze mit wenig Objekten eventuell nicht mehr erfüllt, sodass eine verlässliche Schätzung der Regressionskoeffizienten nur bei Datensätzen mit einer hinreichend großen Objektanzahl gewährleistet ist.

Tabelle 3 zeigt für die größte untersuchte Ausfallquote $q = 25\ \%$ eine Übersicht zum Anteil und der Anzahl der vollständigen Objekte der Datensätze. Die angegebenen Werte stellen die (gerundeten) Erwartungswerte der vollständigen Objekte und Objekte pro zu schätzenden Regressionskoeffizienten dar, wenn die fehlenden Werte zufällig und unabhängig voneinander in den vom Ausfall betroffenen Merkmalen auftreten. Neben dem Hittersdatensatz, für den keine eindeutige Schätzung der Regressionskoeffizienten gewährleistet ist, weisen auch der Auto- und Krebsdatensatz mit 14 bzw. 12 weniger als die von Bankhofer und Vogel (2008, S. 228) empfohlenen 20 Objekte pro zu schätzenden Regressionskoeffizienten auf. Unter der Ausfallquote $q = 25\ \%$ erzielt die CCA bei den Bewertungskriterien relative Abweichung der Regressionskoeffizienten und Abweichung der prognostizierten Werte der abhängigen Variable beim Autodatensatz mäßige bis schlechte Ergebnisse und beim Krebsdatensatz mit Abstand die schlechtesten Ergebnisse. Bei den Datensätzen, bei denen 74 oder mehr Objekte pro zu schätzenden Regressionskoeffizienten zur Verfügung stehen, erreicht

die CCA hingegen gute bis sehr gute Ergebnisse. Die CCA profitiert somit, bei ähnlicher Merkmalsanzahl, von einer zunehmenden Objektanzahl.

	Auto	**Irismuschel**	**Kalifornien-Hauspreis**	**Hitters**	**Krebs**	**Energie**
Objektanzahl	392	4177	20433	263	3047	19735
Merkmalsanzahl	8	10	9	20	18	19
Anteil vollständiger Objekte ($q = 25$ %)	31,64 %	23,73 %	31,64 %	5,63 %	7,51 %	7,51 %
Anzahl vollständiger Objekte ($q = 25$ %)	124	991	6465	15	229	1482
Anzahl vollständiger Objekte pro zu schätzenden Regressionskoeffizienten ($q = 25$ %)	14	90	647	1	12	74

Tabelle 3: *Vollständige Objekte unter einer Ausfallquote von 25 % (Quelle: Eigene Darstellung)*

Aus Tabelle 3 geht zudem hervor, dass infolge einer höheren **Merkmalsanzahl** auch ein höherer Anteil unvollständiger Objekte vorliegt. Infolgedessen weist die CCA im Rahmen der Simulationsstudie bei einer ähnlichen Objektanzahl Nachteile durch eine höhere Merkmalsanzahl auf.[14] Die notwendige Objektanzahl für eine verlässliche Schätzung der Regressionskoeffizienten nach Anwendung der CCA hängt somit auch von der Anzahl der Merkmale mit fehlenden Werten ab.

Bei den untersuchten Imputationsverfahren ist kein eindeutiger Einfluss der **Merkmalsanzahl** auf alle Imputationsverfahren ersichtlich, sodass erneut eine detailliertere Betrachtung der einzelnen Verfahren und Bewertungskriterien erforderlich ist. Mögliche Ursachen hierfür können Wechselwirkungen zwischen einzelnen Faktoren und den Imputationsverfahren sowie weitere (nicht systematisch variierte) Eigenschaften der realen Datensätze, wie z. B. die Korrelation der Merkmale, sein. Im Folgenden soll daher ein Vergleich der Imputationsverfahren in Relation zueinander erfolgen, um zu überprüfen, wie sich eine Erhöhung der Merkmalsanzahl auf die Ergebnisse auswirkt.

[14] Die Ursache hierfür besteht darin, dass im Rahmen der Simulationsstudie die Anzahl der Merkmale mit fehlenden Werten von der Merkmalsanzahl gesteuert wird (vgl. Kapitel 3).

Während die SRI im Allgemeinen wenig Veränderungen infolge einer erhöhten Merkmalsanzahl aufweist und sich lediglich beim Vergleich von Auto- und Hittersdatensatz bei der Abweichung des korrigierten Bestimmtheitsmaßes und der Regressionskoeffizienten merklich verschlechtert, zeigt die DRI über alle Bewertungskriterien hinweg deutliche Verschlechterungen bei dem Hitters- und Energiedatensatz im Vergleich zum Auto- bzw. Kalifornien-Hauspreisdatensatz. Beim Vergleich zwischen Irismuschel- und Krebsdatensatz ist diese ausgeprägte Verschlechterung der DRI allerdings nicht zu erkennen. Eine mögliche Ursache für die deutliche Verschlechterung beim Hittersdatensatz könnte das geringe Verhältnis zwischen Objekt- und Merkmalsanzahl sein, das mit ca. 13: 1 deutlich niedriger ist als bei den anderen Datensätzen.[15]

Die IMW verbessert sich beim Übergang vom Auto- zum Hittersdatensatz in Relation zu den anderen Imputationsverfahren bei allen Bewertungskriterien. Auch beim Vergleich des Irismuschel- bzw. Krebsdatensatzes weist die IMW bei einer höheren Merkmalsanzahl bessere Ergebnisse bezüglich der Abweichung der Regressionskoeffizienten und der Prognose der Werte der abhängigen Variable auf. Bei den Datensätzen mit sehr vielen Objekten ist kein merklicher Unterschied im Hinblick auf die Merkmalsanzahl ersichtlich.

Bei den Hot-Deck Imputationsverfahren zeigt sich, dass die SRHDI beim Übergang vom Auto- zum Hittersdatensatz zu deutlich ähnlicheren Schätzungen der Regressionskoeffizienten als die anderen Imputationsverfahren führt, aber bei den übrigen Datensätzen und Bewertungskriterien weiterhin schlechte Ergebnisse erreicht. Die NNHDI profitiert bei allen drei Bewertungskriterien von der höheren Merkmalsanzahl des Energiedatensatzes gegenüber dem Kalifornien-Hauspreisdatensatz. Bei den anderen Datensätzen schneidet die NNHDI infolge einer höheren Merkmalsanzahl, bei ähnlicher Objektanzahl, aber fast immer schlechter ab. Die einzige Ausnahme besteht beim Vergleich des Auto- und Hittersdatensatzes bezüglich der Abweichung des korrigierten Bestimmtheitsmaßes. Eine Ursache hierfür könnte darin bestehen, dass durch die höhere Merkmalsanzahl weniger vollständige Objekte als mögliche Spender zur Verfügung stehen (vgl. Tabelle 3) und einzelne Objekte somit sehr häufig als Spender fungieren. Die daraus resultierenden Nachteile können die Vorteile einer höheren Merkmalsanzahl bei der Bestimmung des ähnlichsten Spenders überwiegen, sodass

[15] Das nächst niedrigste Verhältnis zwischen Objekt- und Merkmalsanzahl weist der Autodatensatz mit 49: 1 auf.

die NNHDI nur bei einer hinreichend großen Objektanzahl von einer zunehmenden Merkmalsanzahl profitiert.

Zusammenfassend zeigt sich, dass eine Erhöhung der Ausfallquote zu schlechteren Ergebnissen führt. Auch der Übergang von einem MCAR zu einem MAR oder MNAR Ausfallmechanismus geht tendenziell mit einer Verschlechterung der Ergebnisse einher. Der Einfluss der Objekt- und Merkmalsanzahl ist hingegen nicht eindeutig und hängt von weiteren Eigenschaften des jeweiligen Datensatzes, wie z. B. der Korrelation der Merkmale, ab. Dabei zeigt sich, dass Wechselwirkungen zwischen diesen beiden Einflussgrößen bestehen und einige MD-Verfahren nur bei einer hinreichend großen Objektanzahl von einer Erhöhung der Merkmalsanzahl profitieren und sich bei einer zu geringen Objektanzahl infolge der höheren Merkmalsanzahl verschlechtern. Die Objektanzahl sollte daher stets in Zusammenhang mit der Merkmalsanzahl, bzw. der Anzahl der Merkmale mit fehlenden Werten, betrachtet werden.

Des Weiteren variieren die Ergebnisse teilweise deutlich zwischen den Bewertungskriterien, sodass es nicht möglich ist, mit Hilfe der Ergebnisse eines Bewertungskriteriums unmittelbar auf die Ergebnisse eines anderen Bewertungskriteriums zu schließen. Die Frage, welches MD-Verfahren zu der ähnlichsten Regressionsfunktion bzw. den ähnlichsten Ergebnissen führt, kann daher nicht für alle Situationen einheitlich beantwortet werden. Besonderes Augenmerk sollte auf die teilweise sehr unterschiedlichen Ergebnisse zwischen der Abweichung der Regressionskoeffizienten und den prognostizierten Werten der abhängigen Variable gelegt werden. Die Ergebnisse in Kapitel 4 zeigen, dass eine Regressionsfunktion mit stark abweichenden Regressionskoeffizienten zu deutlich ähnlicheren Prognosen der abhängigen Variable führen kann als eine Regressionsfunktion mit ähnlichen Regressionskoeffizienten. Dabei zeigt sich, dass insbesondere bei einem geringen Verhältnis zwischen Objekt- und Merkmalsanzahl, wie z. B. beim Hittersdatensatz, sehr unterschiedliche Ergebnisse resultieren können. Die zweckmäßige Beurteilung, welches der MD-Verfahren zu der ähnlichsten Regressionsfunktion führt, hängt somit auch von der jeweiligen Zielstellung bzw. der späteren Verwendung der Regressionsfunktion ab. Dabei kann beispielsweise zwischen einer Prognose der abhängigen Variable sowie einer Zusammenhangsanalyse und unmittelbaren Interpretation der Regressionskoeffizienten unterschieden werden.

6 Fazit und Ausblick

Ziel dieses Arbeitspapiers war es, die Auswirkungen von fehlenden Daten auf die Ergebnisse der multiplen linearen Regression zu untersuchen und zu ermitteln, unter welchen Bedingungen fehlende Daten zu einer ähnlichen bzw. einer sehr verschiedenen Regressionsfunktion im Vergleich zu vollständigen Daten führen. Die gezielte Untersuchung der Auswirkungen verschiedener Einflussgrößen auf die Ergebnisse der multiplen linearen Regression erfolgte im Rahmen einer Simulationsstudie. Anhand von sechs realen Datensätzen und sechs MD-Verfahren wurde die Variation von Objektanzahl, Merkmalsanzahl, fünf Ausfallquoten und drei Ausfallmechanismen untersucht. Dabei erfolgte die Beurteilung der Auswirkungen der fehlenden Daten anhand der drei Bewertungskriterien: relative Abweichung der Regressionskoeffizienten, Abweichung des korrigierten Bestimmtheitsmaßes und Abweichung der prognostizierten Werte der abhängigen Variable.

Die Simulationsergebnisse zeigten, dass keines der untersuchten MD-Verfahren allen anderen Verfahren in jeder Hinsicht überlegen ist. Des Weiteren variierten die Ergebnisse teilweise deutlich zwischen den Bewertungskriterien, sodass es nicht möglich war, mit Hilfe der Ergebnisse eines Bewertungskriteriums, unmittelbar auf die Ergebnisse eines anderen Bewertungskriteriums zu schließen. Besonderes Augenmerk sollte auf die teilweise unterschiedlichen Ergebnisse zwischen der Abweichung der Regressionskoeffizienten und den prognostizierten Werten der abhängigen Variable gelegt werden. Die zweckmäßige Beurteilung, welches der MD-Verfahren zu der ähnlichsten Regressionsfunktion führt, hängt somit auch von der jeweiligen Zielstellung bzw. der späteren Verwendung der Regressionsfunktion ab. Dabei wurde beispielsweise zwischen einer Prognose der abhängigen Variable sowie einer Zusammenhangsanalyse und unmittelbaren Interpretation der Regressionskoeffizienten unterschieden. Die jeweils besten Verfahren der einzelnen Datensätze sind in Tabelle 4 zusammengefasst. Aus Tabelle 4 geht hervor, dass die SRI, gefolgt von der NNHDI, über alle Bewertungskriterien und/oder Datensätze hinweg oftmals zu den besten Ergebnissen führte. Die übrigen Verfahren konnten nur vereinzelt die besten Ergebnisse erreichen.

	Auto	**Irismuschel**	**Kalifornien-Hauspreis**	**Hitters**	**Krebs**	**Energie**
Objektanzahl	392	4177	20433	263	3047	19735
Merkmalsanzahl	8	10	9	20	18	19
Regressionskoeffizienten	SRI NNHDI	SRI NNHDI	SRI	SRI ($q \leq 5\,\%$) SRHDI ($q \geq 10\,\%$)	SRI DRI ($q \leq 10\,\%$)	SRI
Korrigiertes Bestimmtheitsmaß	SRI	SRI	SRI	SRI ($q \leq 10\,\%$) NNHDI ($q \geq 15\,\%$)	SRI	SRI
Prognostizierte Werte	SRI NNHDI	SRI NNHDI ($q \leq 20\,\%$)	SRI CCA (MAR Ausfallmechanismus nur für $q \leq 10\,\%$)	SRI	SRI	SRI NNHDI ($q \leq 10\%$)

Tabelle 4: *Beste MD-Verfahren der einzelnen Datensätze (Quelle: Eigene Darstellung)*

Die Auswirkungen verschiedener Faktoren auf die Ergebnisse kann folgendermaßen zusammengefasst werden: Eine höhere Ausfallquote sowie der Übergang von einem MCAR zu einem MAR oder MNAR Ausfallmechanismus führen zu größeren Abweichungen der Regressionsfunktionen. Einige MD-Verfahren profitieren von einer höheren Objektanzahl und führen infolgedessen zu besseren Ergebnissen. Der Einfluss der Merkmalsanzahl hängt insbesondere von der Objektanzahl, aber auch von den weiteren Eigenschaften des Datensatzes sowie dem gewählten MD-Verfahren, ab. Im Allgemeinen sollten stets Wechselwirkungen zwischen den verschiedenen Einflussgrößen und MD-Verfahren sowie der spätere Verwendungszweck der Regressionsfunktion berücksichtigt werden, um beurteilen zu können, welches Vorgehen die ähnlichste Regressionsfunktion verspricht.

Da im Rahmen dieser Arbeit ausschließlich reale Datensätze verwendet wurden, besteht weiteres Forschungspotenzial hinsichtlich Einflussgrößen, die unter Verwendung von realen Datensätzen akzeptiert werden mussten. Von besonderem Interesse könnte dabei die Korrelation der Merkmale und die resultierenden Wechselwirkungen mit anderen Einflussgrößen sein. Diese Untersuchungen könnten auch anhand von simulierten Datensätzen erfolgen, da

hierdurch alle Eigenschaften der Datensätze in der Simulation direkt variiert werden können. Die Verwendung von realen Datensätzen ermöglichte allerdings die Untersuchung verschiedener Ausfallmechanismen und MD-Verfahren anhand von „echten“ Daten, sodass vermehrt Wechselwirkungen zwischen Einflussgrößen auftreten und betrachtet werden konnten, die unter Verwendung von simulierten Datensätzen eventuell nicht variieren würden. Beide Vorgehensweisen stellen somit berechtigte Optionen zur Untersuchung der Auswirkungen fehlender Daten dar.

Ein anderes Ziel zukünftiger Forschungsvorhaben könnte in der detaillierten Untersuchung der notwendigen Objektanzahl sowie des maximalen Anteils unvollständiger Objekte für eine zufriedenstellende Anwendung der CCA bestehen. Auch die systematische Untersuchung der Vor- und Nachteile einer höheren Merkmalsanzahl bei Anwendung der NNHDI stellt ein mögliches Gebiet zukünftiger Forschung dar.

Literaturverzeichnis

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.

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

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

Bankhofer, Udo; Vogel, Jürgen (2008): Datenanalyse und Statistik. Eine Einführung für Ökonomen im Bachelor. Wiesbaden: Gabler. Online verfügbar unter http://dx.doi.org/10.1007/978-3-8349-9654-1.

Bartlett, Jonathan W.; Carpenter, James R.; Tilling, Kate; Vansteelandt, Stijn (2014): Improving upon the efficiency of complete case analysis when covariates are MNAR. In: *Biostatistics* 15 (4), S. 719–730. DOI: 10.1093/biostatistics/kxu023.

Bartlett, Jonathan W.; Harel, Ofer; Carpenter, James R. (2015): Asymptotically Unbiased Estimation of Exposure Odds Ratios in Complete Records Logistic Regression. In: *American Journal of Epidemiology* 182 (8), S. 730–736. DOI: 10.1093/aje/kwv114.

Demissie, Serkalem; LaValley, Michael P.; Horton, Nicholas J.; Glynn, Robert J.; Cupples, L. Adrienne (2003): Bias due to missing exposure data using complete-case analysis in the proportional hazards regression model. In: *Statistics in Medicine* 22 (4), S. 545–557. DOI: 10.1002/sim.1340.

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

Enders, Craig K. (2010): Applied missing data analysis. New York: The Guilford Press. Online verfügbar unter http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10389908.

Fahrmeir, Ludwig; Kneib, Thomas; Lang, Stefan (2009): Regression. Modelle, Methoden und Anwendungen. 2. Aufl. Berlin, Heidelberg: Springer (Statistik und ihre Anwendungen). Online verfügbar unter http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10328772.

Ford, B. L. (1983): An Overview of Hot-Deck Procedures. In: W. G. Madow, H. Nisselson und I. Olkin (Hg.): Incomplete Data in Sample Surveys. Volume 2: Theory and Bibliographies. New York: Academic Press, S. 185–207.

Frane, James W. (1976): Some simple procedures for handling missing data in multivariate analysis. In: *Psychometrika* 41 (3), S. 409–415. DOI: 10.1007/BF02293565.

Glynn, Robert J.; Laird, Nan M. (1986): Regression Estimates and Missing Data: Complete Case Analysis. Harvard School of Public Health, Department of Biostatistics.

James, Gareth; Witten, Daniela; Hastie, Trevor; Tibshirani, Robert (2013): An Introduction to Statistical Learning. with Applications in R. New York, Heidelberg, Dordrecht, London: Springer.

Joenssen, Dieter William (2015a): HotDeckImputation. Hot Deck Imputation Methods for Missing Data. Version 1.1.0. Online verfügbar unter https://CRAN.R-project.org/package=HotDeckImputation.

Joenssen, Dieter William Hermann (2015b): Hot-Deck-Verfahren zur Imputation fehlender Daten. Auswirkungen des Donor-Limits. Universitätsbibliothek Ilmenau, Ilmenau.

Joenssen, Dieter William Hermann; Bankhofer, Udo (2012): Hot Deck Methods for Imputing Missing Data. The Effects of Limiting Donor Usage. In: Petra Perner (Hg.): Machine Learning and Data Mining in Pattern Recognition. Berlin, Heidelberg: Springer, S. 63–75.

Little, Roderick J. A. (1992): Regression With Missing X's: A Review. In: *Journal of the American Statistical Association* 87 (420), S. 1227–1237. DOI: 10.2307/2290664.

Little, Roderick J. A.; Rubin, Donald B. (2020): Statistical Analysis with Missing Data. 3. Aufl. Hoboken: John Wiley & Sons.

Pace, R. Kelley; Barry, Ronald (1997): Sparse spatial autoregressions. In: *Statistics & Probability Letters* 33 (3), S. 291–297. DOI: 10.1016/S0167-7152(96)00140-X.

Peugh, James L.; Enders, Craig K. (2004): Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement. In: *Review of Educational Research* 74 (4), S. 525–556. DOI: 10.3102/00346543074004525.

R Core Team (2020): R. A Language and Environment for Statistical Computing. Version 3.6.3. Vienna, Austria. Online verfügbar unter https://www.R-project.org/.

Rockel, Tobias (2017): Gütevergleich von Imputationsverfahren - Eine Analyse existierender Simulationsstudien. Ilmenau: Universitätsbibliothek Ilmenau (Ilmenauer Beiträge zur Wirtschaftsinformatik). Online verfügbar unter https://nbn-resolving.org/urn:nbn:de:gbv:ilm1-2017200274.

Rockel, Tobias (2018): Vergleich von Imputationsverfahren - Eine Simulationsstudie. Ilmenau: Universitätsbibliothek Ilmenau (Ilmenauer Beiträge zur Wirtschaftsinformatik). Online verfügbar unter https://nbn-resolving.org/urn:nbn:de:gbv:ilm1-2018200160.

RStudio Team (2020): RStudio. Integrated Development Environment for R. Boston, MA. Online verfügbar unter http://www.rstudio.com/.

Rubin, Donald B. (1976): Inference and missing data. In: *Biometrika* 63 (3), S. 581–592. DOI: 10.2307/2335739.

Sande, I. G. (1983): Hot-Deck Imputation Procedures. In: W. G. Madow, H. Nisselson und I. Olkin (Hg.): Incomplete Data in Sample Surveys. Volume 3: Proceedings of the Symposium. New York: Academic Press, S. 339–349.

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

Schnell, Rainer (1986): Missing-Data-Probleme in der empirischen Sozialforschung. Ruhr-Universität Bochum: Dissertation. Online verfügbar unter http://nbn-resolving.de/urn:nbn:de:bsz:352-opus-5490.

van Buuren, Stef (2018): Flexible Imputation of Missing Data. 2. Aufl. Boca Raton, London, New York: CRC Press. Online verfügbar unter https://www.taylorfrancis.com/books/9780429960352.

van Buuren, Stef; Groothuis-Oudshoorn, Karin (2011): mice: Multivariate Imputation by Chained Equations in R. In: *Journal of Statistical Software* 45 (3). DOI: 10.18637/jss.v045.i03.