Mikrozensus

Maschinelles Lernen: *Classification and Regression Trees (CART)* für die Imputation nutzbar machen

von **Birgit Pech**

Wenn Imputationsmodelle aufgrund unzutreffender Vorannahmen nicht angemessen spezifiziert werden, führt dies in anschließenden Analysen zu verzerrten Schätzern. Mit dem Ziel, Imputationsverfahren bereitzustellen, die komplexe Datenstrukturen im Datensatz adaptiv erfassen können und dadurch Fehlspezifikationen vermeiden helfen, werden zunehmend Methoden des maschinellen Lernens eingesetzt. Dieser Beitrag fokussiert dabei auf Classification and Regression Trees (CART). Nach einer Einführung in die Methode wird ein Anwendungsbeispiel mit Daten des Mikrozensus 2016 vorgestellt, für den fehlende Werte in der Variable „Anzahl geborener Kinder" ergänzt wurden. Hierbei wurde ein Methodenvergleich zwischen einem CART-basierten Imputationsverfahren und dem Predictive Mean Matching durchgeführt.

1. Einleitung

Werden Imputationsmodelle zur modellgestützten Ergänzung fehlender Werte nicht angemessen spezifiziert, führt dies in anschließenden Analysen zu verzerrten Schätzern. So kann in einem Imputationsmodell fälschlicherweise von Linearität in den Modellparametern ausgegangen werden, während tatsächlich nichtlineare Strukturen vorherrschen, oder das Modell geht von additiven Effekten aus, während tatsächlich komplexe Interaktionseffekte den datengenerierenden Prozess charakterisieren.

Mit dem Ziel, Imputationsverfahren bereitzustellen, die komplexe Datenstrukturen im Datensatz adaptiv erfassen können, werden auch in der amtlichen Statistik vermehrt Methoden des maschinellen Lernens eingesetzt. Das hier vorgestellte CART-Verfahren ist ein nichtparametrisches Verfahren, das für die Ergänzung kategorialer wie auch metrischer Variablen nutzbar gemacht werden kann. Im ersten Fall handelt es sich um *Classification Trees*, im zweiten um *Regression Trees*.

Wie bei allen nichtparametrischen Verfahren müssen keine Vorannahmen über die bedingte Verteilung der im Modellzusammenhang abhängigen Variablen getroffen werden, wodurch es flexibler einsetzbar ist. Darüber hinaus ist beim CART-Verfahren weder die a priori-Spezifikation der Modellvariablen noch eine Spezifikation der konkreten Modellbeziehungen, beispielsweise von Interaktionsbeziehungen zwischen den Prädiktoren, notwendig. Die entsprechenden Datenstrukturen werden vielmehr algorithmisch identifiziert. Gleichwohl bleibt eine Vorselektion von potenziell besonders einflussreichen Modellvariablen für die weitere algorithmische Prüfung möglich, beispielsweise zur Effizienzsteigerung der Berechnung.

Nach einer Einführung in die Methode wird ein Anwendungsbeispiel mit Daten des Mikrozensus 2016 vorgestellt, in welchem fehlende Werte für das Merkmal „Anzahl geborener Kinder" ergänzt werden. Dabei wird ein Methodenvergleich zwischen einem CART-basierten Imputationsverfahren und dem Predictive Mean Matching (PMM) durchgeführt. Letzteres ist die aktuell vom Statistischen Bundesamt (Destatis) genutzte Methode zur Ergänzung fehlender Werte in der Variable „Anzahl geborener Kinder" im Mikrozensus.

2. Einführung in die CART-Methode

CART-Algorithmen sind ursprünglich im Data Mining-Kontext entstanden (vgl. Breiman et al. 1984). Primäres Ziel war die möglichst fehlerfreie Fallvorhersage oder -klassifikation. Eine solche Analyse dient beispielsweise dazu, aus gegebenen individuellen Hilfsinformationen abzuleiten, für welche Produktgruppe sich ein potenzieller Käufer interessiert, um das jeweilige Produkt dann gezielt zu bewerben. Das Vorhersagemodell wird anhand eines Lerndatensatzes mit vollständigen Informationen über die Produktpräferenzen entwickelt, um es dann auf Fälle mit noch unbekannten Produktpräferenzen zu übertragen. Ein geeignetes Qualitätskriterium für Analysen solcher Art ist die Treffer- bzw. Fehlerquote der richtig vs. falsch klassifizierten Fälle.

Zielsetzung bei der Imputation ist hingegen nicht die fallbezogen möglichst fehlerfreie Vorhersage oder Rekonstruktion von Daten, sondern die Wiederherstellung von Verteilungsstrukturen: die Daten sollen so ergänzt werden, dass die Verteilungsstrukturen rekonstruiert werden, wie sie sich ohne fehlende Werte darstellen würden. Auf der Basis der vervollständigten Daten ist dann die Berechnung

verzerrungsfreier und präziser Schätzer möglich. Qualitätskriterium ist hier also die Verzerrungsfreiheit und Präzision von Schätzgrößen, beispielsweise von Mittelwert-, Anteilswert- oder Regressionsschätzern, auf Basis der vervollständigten Daten.

Die beiden Zielsetzungen – geringe Fehlerquote bei Fallvorhersagen vs. Rekonstruktion von Verteilungsstrukturen – sind nicht deckungsgleich. Daher werden CART-Algorithmen, die ursprünglich für ersteren Zweck entwickelt wurden, für Imputationszwecke angepasst, indem Zufallsprozesse eine größere Bedeutung erhalten. Auch vergleichsweise unwahrscheinliche Werte bekommen damit eine angemessene Chance, im gesamten Datensatz zumindest selten imputiert zu werden, und Zufallsprozesse sorgen dafür, dass sich gegebene Schätzunsicherheiten in der Varianz der imputierten Werte widerspiegeln.

Für den Zweck möglichst fehlerfreier individueller Fallvorhersagen stehen inzwischen fortgeschrittenere Machine Learning-Algorithmen bereit. Für Zwecke der Imputation haben sich in Methodenvergleichen gleichwohl die älteren, CART-basierten Algorithmen als die vielversprechendere Alternative erwiesen (vgl. beispielsweise Reiter 2005, Burgette und Reiter 2010, Drechsler und Reiter 2011, Doove et al. 2014, Loh et al. 2019).

Der Imputationsprozess erfolgt in zwei Schritten, die im weiteren Verlauf ausführlicher erläutert werden (vgl. Berk 2008, Hastie et al. 2008):

- Im ersten Schritt (Modellschritt) wird der Datensatz mit den vollständigen Fällen sukzessive aufgesplittet. Das Endergebnis lässt sich als Entscheidungsbaum mit der „Wurzel" oben, den „Blättern" oder Endknoten (englisch terminal nodes) unten darstellen. Die Responsewerte der Fälle in den Endknoten dienen im zweiten Schritt als Kandidaten für die Imputationswerte.
- Im zweiten Schritt (Imputationsschritt) werden die unvollständigen Fälle dann gemäß den Entscheidungsbaum-Regeln ihrem Endknoten zugewiesen. Aus diesen werden so viele Imputationswerte wie nötig zufällig gezogen.

Abbildung a soll zunächst dem Verständnis des Modellschritts dienen. In diesem Beispiel ist Y eine kategoriale Variable mit vier Ausprägungen (hier farblich unterschieden); X1 und X2 sind zwei metrische Einflussvariablen (Prädiktoren). Der Scatterplot links ergibt sich aus den vorliegenden Daten, die analysiert werden sollen. Der Entscheidungsbaum rechts zeigt die Datenstruktur, die der CART-Algorithmus adaptiv identifiziert.[1] Dabei geht der Algorithmus wie im Folgenden beschrieben vor.

Zur Modellbildung splittet der CART-Algorithmus den Lerndatensatz mit den vollständigen Beobachtungen sukzessive binär auf. Jeder Split wird so gewählt, dass möglichst homogene Teildatensätze hinsichtlich der Ausprägungen der Responsevariable entstehen. Dazu wird jeder Prädiktor und jeder mögliche binäre Split der Prädiktorwerte[2] überprüft und der Split mit der besten Homogenisierungswirkung ausgewählt.

Gesplittet wird, wenn durch den Split Heterogenität reduziert wird[3]. Die Reduktion von Heterogenität durch einen potenziellen Split s eines Elternknotens A ist definiert als die Differenz zwischen der Heterogenität des Elternknotens H_A minus den gewichteten Heterogenitäten der potenziellen linken und rechten Tochterknoten. Die Gewichtung bemisst sich am Anteil der Fälle $\hat{p}$ in den potenziellen Tochterknoten:

$$\Delta H_{s,A} = H_A - (H_{\text{TochterA\_li}} * \hat{p}_{\text{TochterA\_li}} + H_{\text{TochterA\_re}} * \hat{p}_{\text{TochterA\_re}})$$

Das Heterogenitätsmaß für metrische Responsevariablen (im Falle von *Regression Trees*) ist die Streuung gemessen an der Summe der quadrierten Abweichungen vom Mittelwert im betrachteten Eltern- bzw. Tochterknoten:

$$SS = \sum_{i=1}^{n\_node} (y_i - \bar{y}_{node})^2$$

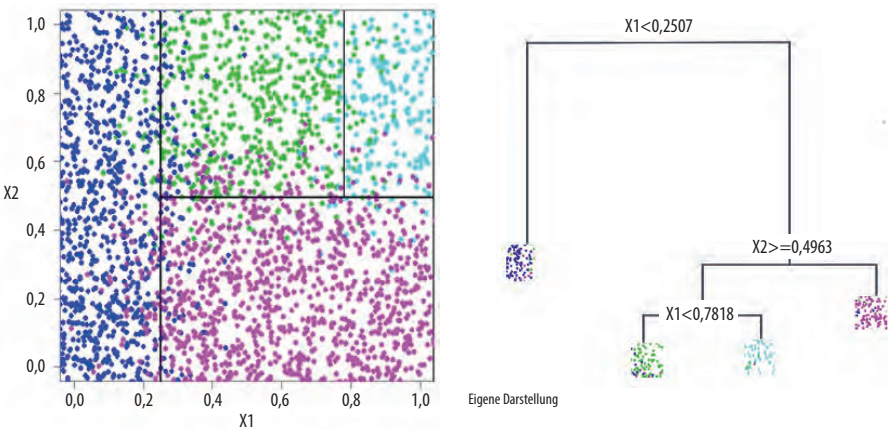
Das Heterogenitätsmaß für kategoriale Responsevariablen (im Falle von *Classification Trees*) ist in der Regel[4] das kategoriale Gini-Maß:

$$Gini = \sum_{k=1}^{K\_node} \hat{p}_k * (1 - \hat{p}_k)$$

mit $\hat{p}_k$ als relativer Häufigkeit einer Kategorie k. Beide Maße haben Null als Minimalwert.

Mit der Aufsplittung des Datensatzes (Partitionierung) wird weiterverfahren, bis in den Endknoten entweder nicht mehr zu verbessernde Homogenität

a | Funktionsweise des CART-Algorithmus



Eigene Darstellung

1 Die Splitformulierung in der Grafikbeschriftung bezieht sich immer auf den linken Tochterknoten.

2 Bei kategorialen Prädiktoren wird jede Splitmöglichkeit ohne Beachtung der Reihenfolge überprüft. Beispielsweise kann die Verästelung bei einer kategorialen Prädiktorvariable mit vier Ausprägungen {1, 2, 3, 4} so gestaltet sein, dass die Ausprägungen {1, 3} in den linken und die Ausprägungen {2, 4} in den rechten Tochterknoten weitergeleitet werden. Bei metrischen Prädiktoren wird dagegen die Reihenfolge beachtet. Ein Tochterknoten enthält immer Werte kleiner, der andere größer als oder gleich dem Splitwert.

3 Im Sinne eines Abbruchkriteriums kann auch festgelegt werden, dass diese Reduktion eine bestimmte Mindestgröße annehmen muss.

4 Dies ist die Voreinstellung vieler Softwarepakete, beispielsweise in *mice* (van Buuren und Groothuis-Oudshoorn 2019) und *rpart* (Therneau und Atkinson 2019).

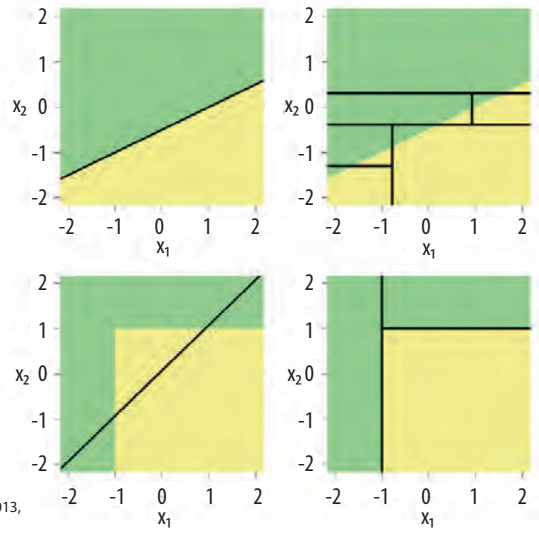
erreicht ist, oder bis ein Abbruchkriterium erfüllt ist, z. B. dass der Endknoten mindestens 10 Fälle enthalten muss. Sinn von Abbruchkriterien ist es, ein *Overfitting* zu vermeiden, also den Einfluss spezifischer Besonderheiten im Lerndatensatz zu begrenzen und die Generalisierbarkeit des Modells zu erhöhen.[5] Die in den Endknoten beobachteten Werte dienen im zweiten Schritt als Kandidaten für die Imputationswerte.

Im zweiten Schritt (Imputationsschritt) wird das Entscheidungsbaummodell auf neue Fälle angewendet. Diese Fälle haben fehlende Werte in der Responsevariable Y, die ergänzt, also imputiert werden sollen. Bekannt seien aber die Ausprägungen der Prädiktorvariablen (im Beispiel X1 und X2). Je nachdem, welche Werte die Prädiktorvariablen im jeweiligen Einzelfall annehmen, werden diese Fälle ihrem zugehörigen Endknoten zugewiesen: Die unvollständigen Fälle wandern – metaphorisch ausgedrückt – an den Ästen des Entscheidungsbaums entlang bis zu ihrem Endknoten. Aus den entsprechenden Endknoten werden schließlich so viele Imputationswerte wie nötig zufällig (mit Zurücklegen) gezogen.[6]

Zu klären ist noch, wie mit unvollständigen Prädiktorvariablen umgegangen wird, denn die Prädiktoren werden ja einerseits für die Modellgenerierung, andererseits für die Zuweisung der unvollständigen Fälle zu ihren Endknoten benötigt. Fehlen die Werte in einer oder mehreren Prädiktorvariablen und fungieren diese als Splitvariablen, droht der betreffende Fall an der betreffenden Verästelung im Baum „steckenzubleiben", da unklar wäre, wie mit diesem Fall weiter zu verfahren ist. Er könnte also nicht zu seinem Endknoten weitergeleitet werden.

Einige Softwarepakete schließen daher Fälle mit unvollständigen Prädiktorvariablen aus dem Imputationsprozess aus, es sei denn, die Prädiktorvariablen wurden zuvor selbst vervollständigt. Dies ist beispielsweise für das häufig für Imputationsprojekte genutzte R-Programm *mice* der Fall, welches neben vielen anderen Methoden auch eine CART-basierte Imputationsmethode anbietet.

b | **CART und Linearität**



aus:
James et al. 2013, S. 314

Ein R-Paket, das spezifischer auf den CART-Algorithmus zugeschnitten ist – allerdings im Zuge der Programmierung zuerst angepasst werden muss, wenn CART statt für Fallvorhersagen für die Imputation nutzbar gemacht werden soll – ist das R-Paket *rpart* (Therneau und Atkinson 2019). Als CART-spezifischeres Paket bietet es eine Reihe von nützlichen Spezialfunktionen an, die auch mit Blick auf Imputationsvorhaben hilfreich sind, darunter verschiedene Varianten des Umgangs mit fehlenden Prädiktorvariablen. So ist es mit *rpart* unter anderem möglich, Ersatzvariablen bzw. Ersatzsplits mit ähnlicher Homogenisierungswirkung zu identifizieren, wie dies für den Originalsplit der Fall gewesen wäre (die Ersatzvariablen werden „Surrogatvariablen", die Ersatzsplits „Surrogatsplits" genannt). Droht ein Fall wegen fehlender Prädiktorwerte „steckenzubleiben", gibt der Surrogatsplit vor, wie der betreffende Fall weitergeleitet wird.[7]

Aus der Literatur zur Nutzung von CART als Imputationsvehikel lassen sich als wichtigste Vor- und Nachteile die folgenden benennen (vgl. Reiter 2005, Burgette and Reiter 2010, Drechsler und Reiter 2011, Doove et al. 2014, Loh 2014, Loh et al. 2019): Das Verfahren hilft, Fehlspezifikationen durch unzureichende Modellspezifikationen zu vermeiden. Der Algorithmus ist insbesondere gut geeignet, um Interaktionseffekte (auch solche höherer Ordnung) und nichtlineare Zusammenhänge adaptiv zu ermitteln, welche besonders schwer a priori zu spezifizieren sind. Ein weiterer Vorteil ist, dass Multikollinearität für die Anwendung des CART-Algorithmus unproblematisch ist, anders als beispielsweise für regressionsbasierte parametrische Modelle. Für die Imputation mit CART kann Multikollinearität sogar zum Vorteil genutzt werden, weil sich hoch korrelierende Prädiktoren besonders gut als Surrogatvariablen eignen.

Wenn im Datensatz jedoch lineare und additive Strukturen vorherrschen, sind lineare Imputationsmodelle, soweit sie korrekt spezifiziert sind, dem CART-Algorithmus überlegen, da solche Strukturen durch die CART-Partitionierungen weniger gut abgebildet werden können. Abbildung b soll dies illustrieren: In der oberen Bildhälfte wird der Datensatz durch Linearbeziehungen charakterisiert, für die ein richtig spezifiziertes Linearmodell die bessere Alternative darstellt. CART-Modelle können sich nur unzureichend an diese Datenstruktur annähern. In der unteren Bildhälfte ist die umgekehrte Situation skizziert. Im Zweifel kann es hilfreich sein, in Simulationen mit den interessierenden Daten zunächst Methodenvergleiche anzustellen.

5 Die Imputationsliteratur empfiehlt für die Modellbildung eher milde Abbruchkriterien (große Bäume mit homogeneren Endknoten), um zwar einerseits *Overfitting* zu vermeiden, andererseits aber auch die Verzerrungen gering zu halten (vgl. z. B. Doove et al. 2014). Dies spiegelt sich auch in den Voreinstellungen von Imputationssoftware wie *mice* wider.

6 Hierbei wird implizit ein *Missing At Random* (MAR) Fehlermechanismus unterstellt. Dies bedeutet, dass die Ausfallwahrscheinlichkeit innerhalb einer Variable von den Werten der Prädiktoren abhängt, jedoch – zumindest sobald auf diese Prädiktoren bedingt wird – nicht von den fehlenden Werten der Variable selbst.

7 Zwar greift auch *mice* für die Funktion *mice.impute.cart* auf *rpart* zurück, davor werden jedoch die Standardschritte von *mice* durchlaufen, in deren Zuge die Fälle mit fehlenden Prädiktorwerten bereits aus dem Datensatz aussortiert werden. Das Gleiche gilt für weitere vorbereitende Schritte in *mice* wie die Dichotomisierung von kategorialen Prädiktorvariablen zu Dummyvariablen, noch bevor die integrierte *rpart*-Funktion greift. Vergleiche dazu ausführlicher Abschnitt 4.

3. Imputation fehlender Werte zur Anzahl geborener Kinder im Mikrozensus

Der Mikrozensus ist die größte jährliche Haushaltsbefragung der amtlichen Statistik in Deutschland. Fast 1 % aller Haushalte wird zu wirtschaftlichen und sozialen Themen und zum Arbeitsmarkt befragt. Das Mikrozensusgesetz schreibt für die meisten Fragen die Auskunftspflicht vor, sodass der Anteil fehlender Werte im Vergleich zu freiwilligen Erhebungen sehr gering ist.

Der Mikrozensus enthält jedoch auch Fragen zur freiwilligen Beantwortung. Darunter finden sich auch die Fragen zur Mutterschaft („Haben Sie Kinder geboren?") und ggf. zur Anzahl der geborenen Kinder, welche sich an weibliche Befragte im Alter von 15 bis 75 Jahren richten. Beide Fragen werden nicht in jedem Jahr gestellt, sondern wurden bisher nur in den Jahren 2008, 2012, 2016 und 2018 erhoben und sind danach fortlaufend alle vier Jahre vorgesehen. Aufgrund der Freiwilligkeit der Auskunft ist der Anteil fehlender Werte höher als im Mikrozensus üblich. Im hier untersuchten Jahr 2016[8] betrug die Ausfallrate etwa 9 %.

Datenanalysen zeigen, dass die Ausfälle nicht rein zufällig über den Datensatz verteilt sind, sondern dass sie für bestimmte Personengruppen höher sind als für andere. Jüngere, gut gebildete und alleinlebende Frauen haben höhere Ausfallraten bei den freiwilligen Fragen zur Mutterschaft bzw. zur Anzahl der geborenen Kinder. Soweit die Angaben im Gesamtdatensatz vorliegen, ist diese Personengruppe zugleich überproportional häufig kinderlos. Daher steht zu vermuten, dass die Ausfälle die Ergebnisse verzerren, also dass ohne die Ausfälle ein höherer Anteil an Kinderlosen gemessen würde (vgl. Spies und Lange 2018). Um auf Grundlage der Mikrozensus-Stichprobe valide Schlüsse auf die Gesamtbevölkerung ziehen zu können, werden die fehlenden Werte durch das Statistische Bundesamt per modellgestützter Imputation ergänzt. Ob und wie viele Kinder geboren wurden, fasst Destatis dazu in einer Zählvariable mit Ausprägungen von „0" bis „15 und mehr" Kindern zusammen.[9]

Destatis nutzt für die Imputation der fehlenden Werte in der Variable „Anzahl geborener Kinder" die Imputationsmethode *Predictive Mean Matching* (PMM, zurückgehend auf Little 1988) mit Hilfe des R-Pakets *mice*.

PMM und CART folgen im Prinzip einem ähnlichen Grundgedanken: Für Fälle mit fehlenden Werten in der Responsevariable werden „Nachbarn" mit vollständigen Werten in der Responsevariable gesucht, also ähnliche Fälle in Bezug auf die Ausprägungen von Prädiktorvariablen, welche ihrerseits mit der Responsevariable zusammenhängen. Aus den in der Erhebung beobachteten Responsevariablenwerten solcher Nachbarn werden dann die Imputationswerte zufällig gezogen. Die Modellbildung beim Predictive Mean Matching geschieht jedoch auf andere Weise als beim CART-Algorithmus (siehe Statistik erklärt).

Statistik erklärt: Predictive Mean Matching (PMM)

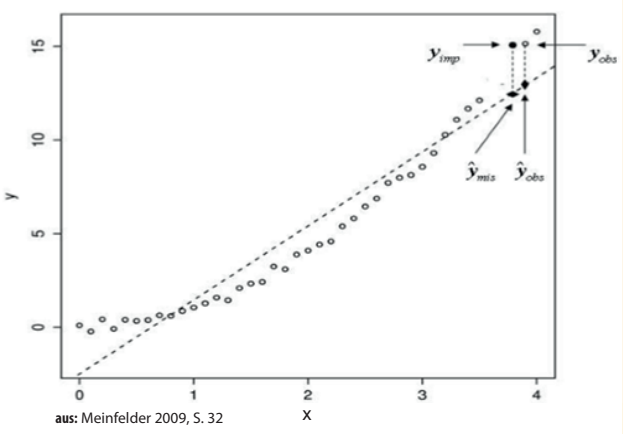
Für die Imputation fehlender Werte in der Variable „Anzahl geborener Kinder" nutzt das Statistische Bundesamt die Methode PMM. Die Imputation fußt auf einem linearen Regressionsmodell. Die Responsevariable Y (hier „Anzahl geborener Kinder") wird auf Prädiktorvariablen X regressiert. Mit Hilfe des Modells lässt sich dann aus gegebenen Werten für X der Wert der Responsevariable $\hat{y}$ (der *predictive mean*) schätzen. Diese Beziehung lässt sich im vereinfachten bivariaten Fall mit Hilfe einer Regressionsgeraden bildlich darstellen (vergleiche Abbildung unten).

Je nachdem, ob es sich um Fälle mit beobachteten Werten in der Responsevariablen handelt oder ob diese Werte fehlen, lässt sich zwischen $\hat{y}_{obs}$ und $\hat{y}_{mis}$ unterscheiden. Ermittelt werden nun für jeden Imputationsfall die fünf $\hat{y}_{obs}$ mit dem geringsten Abstand zum $\hat{y}_{mis}$-Wert (in der Abbildung ist dies der Übersicht halber nur für einen nächsten Nachbarn dargestellt).[1]

Für die Imputation erfolgt dann ein Zufallszug aus den tatsächlich beobachteten y_{obs} dieser fünf nächsten Nachbarn.

Dies macht die Methode zu einer semi-parametrischen Methode: Zwar ist die zugrundeliegende Regressionsschätzung parametrisch, aber es werden nicht die $\hat{y}_{mis}$ selbst imputiert, sondern tatsächlich beobachtete Werte y_{obs}. Dieses Vorgehen hat den Vorteil, dass nur Werte imputiert werden, die in der Realität tatsächlich vorkommen können (anders als z. B. ein exakt geschätzter Wert von 3,8 Kindern). Wie die Abbildung ebenfalls illustriert, ist die Methode dadurch auch robuster gegenüber Abweichungen von der Linearitätsannahme[2], was sie umso vielseitiger anwendbar macht. Den Imputationswert aus einem Pool von *fünf* Nachbarn zu ziehen, statt nur den des nächsten Nachbarn zu nehmen, sorgt für mehr Varianz in den Daten, was der gegebenen Schätzunsicherheit besser entspricht und eine bessere Annäherung an die zu rekonstruierende Ursprungsverteilung ermöglicht.

Predictive Mean Matching



aus: Meinfelder 2009, S. 32

8 Die Daten für das Berichtsjahr 2018 befanden sich zum Zeitpunkt dieser Untersuchung noch in der Aufbereitung durch das Statistische Bundesamt.

9 Auch in der Mutterschaftsvariable gibt es fehlende Werte. Nach der Imputation der Anzahl der geborenen Kinder werden fehlende Fälle in der Variable Mutterschaft deterministisch ergänzt (keine Mutterschaft, wenn der imputierte Wert = 0 ist; Mutterschaft, wenn er > 0 ist). Das Vorgehen, beide Variablen temporär zu einer Zählvariable zusammenzufassen, hatte sich in vorausgehenden methodischen Analysen des Bundesamts als vorteilhaft gegenüber der separaten Imputation beider Variablen erwiesen (vgl. Spies und Lange 2018).

1 Fünf ist der default-Wert in *mice*, welcher auch vom Statistischen Bundesamt und in dieser Vergleichsstudie genutzt wird.

2 PMM erlaubt auch Abweichungen von der Normalverteilungsannahme und der Varianzhomogenität der Residuen (vgl. Gaffert et al. 2016).

Für das zur Schätzung der Anzahl geborener Kinder im Mikrozensus verwendete lineare Regressionsmodell nutzt das Statistische Bundesamt 22 Prädiktorvariablen (siehe Übersicht). Diese Einflussvariablen gehen additiv, also ohne Spezifikation von Interaktionseffekten, in das Destatis-Modell ein. Kategoriale Prädiktoren mit einer großen Anzahl von Kategorien erhalten zumeist zusammenfassende Kodierungen. Ausgefilterte Fälle (beispielsweise Nichterwerbstätige bei der Frage nach Vollzeit-/Teilzeittätigkeit) werden mit einem gültigen Wert versehen (in diesem Beispiel „Sonstige"). Das Bundesamt nutzt zwölf Altersklassen als Imputationsklassen; Modellschritt und Imputationsschritt erfolgen also separat für jede der zwölf Altersklassen, auch um altersspezifischen Besonderheiten Rechnung zu tragen. Gleichwohl geht die Altersvariable als metrische Variable auch innerhalb der Imputationsklassen in die Modellbildung ein (z. B. in der Klasse der 15- bis unter 20-Jährigen mit den Ausprägungen „15" bis „19"). Der Auswahl der Prädiktorvariablen und der Entscheidung für die gewählten Imputationsklassen gingen verschiedene Analysen des Statistischen Bundesamts voraus (vgl. Spies und Lange 2018). Wesentlich waren neben inhaltlichen Erwägungen statistische Zusammenhangsanalysen.

Prädiktorvariablen im PMM-Regressionsmodell[10]

EF7	Erhebungsinstrument
EF44	Alter
EF52	Mutter im Haushalt
EF85	Grund für Nichtarbeit in der letzten Woche
EF92	Grund für Beendigung der letzten Tätigkeit
EF114UG4	Beruf in der gegenwärtigen Tätigkeit
EF117	Gegenwärtige Stellung im Beruf
EF129	Vollzeit/Teilzeittätigkeit
EF130	Grund für Teilzeittätigkeit
EF147	Grund für weniger geleistete Arbeitszeit
EF233	Andere/weitere Tätigkeit gesucht
EF289	Art der besuchten Schule
EF310	Höchster allgemeiner Schulabschluss
EF371	Staatsangehörigkeit
EF439	Bezug von Elterngeld
EF440	Bezug von Kindergeld
EF540	Höchster Grad der allg. oder beruflichen Bildung
EF560U1	Bundesland
EF604	Stadt-Land-Gliederung
EF707	Haushaltsnettoeinkommen
EF765	Familienstand

Wie aufgrund der oben skizzierten Analysen zur Ausfallwahrscheinlichkeit zu erwarten war, stieg nach der Imputation der fehlenden Werte der Anteil der kinderlosen Frauen im Mikrozensus 2016 leicht an.

4. Simulationsstudie zum Methodenvergleich von PMM und CART – Vorgehensweise

Um im gegebenen Anwendungsfall die Güte des von Destatis genutzten Predictive Mean Matching mit der des CART-Imputationsverfahrens zu vergleichen, wurde im Amt für Statistik Berlin-Brandenburg eine Simulationsstudie auf Basis der Mikrozensusdaten 2016 durchgeführt.[11] In den Simulationsdatensatz gingen nur die in der Responsevariable „Anzahl geborener Kinder" vollständigen Fälle des Mikrozensus ein. Mit diesen wurde ein Datenausfall simuliert, also ein Anteil von Fällen künstlich gelöscht. Danach wurden die gelöschten Fälle separat mit Hilfe der zu prüfenden Imputationsmethoden wieder ergänzt. Auf diese Weise konnten verschiedene Verteilungsgrößen (Anteils- und Mittelwerte) vor der Löschung und nach der Imputation miteinander verglichen werden. Der Zyklus des künstlichen Löschens, Imputierens und Berechnens der Analysewerte wurde 200 Mal durchlaufen, um trotz zufälliger Einflüsse bei der Löschung und Imputation der Daten ein zuverlässiges Gesamtbild zu erhalten. Details werden in diesem Abschnitt vorgestellt.

Verglichen wurden die folgenden Imputationsmethoden:

1. PMM.*mice* (Destatis-Imputationsverfahren)
2. CART.*mice*
3. CART.*rpart*
4. CART.*mice.bigmodel*
5. CART.*rpart.bigmodel*

Hinsichtlich des *Predicitive Mean Matching* (PMM.*mice*) wurde genauso vorgegangen wie von Destatis.[12] Zur CART-Methode wurden vier Varianten getestet, die sich in zweifacher Hinsicht unterscheiden: zum einen hinsichtlich des Sets von Prädiktorvariablen (Nutzung der Destatis-Auswahl oder erweitertes Prädiktorvariablenset), zum anderen hinsichtlich der genutzten Software (*mice* oder *rpart*).

Soweit eine der bigmodel-Varianten mit einem erweiterten Prädiktorvariablenset genutzt wurde, kamen zusätzliche Prädiktorvariablen hinzu. Diese wurden aus inhaltlichen Erwägungen heraus gewählt, um zu testen, ob sie die CART-Ergebnisse eventuell noch verbessern könnten. Wie oben erwähnt, spielt Kollinearität für CART keine restringierende Rolle (anders als für die lineare Regression), sodass diesbezügliche Erwägungen bei der Auswahl vernachlässigt werden konnten.

Zusätzliche Prädiktorvariablen im erweiterten Prädiktorvariablenset („bigmodel")

EF120	Überwiegend als Führungs- oder Aufsichtskraft tätig
EF401	Überwiegender Lebensunterhalt
EF770	Zahl der ledigen Kinder in der Familie/Lebensform
EF2009	Migrationsstatus

Die zweite Differenzierung betrifft die genutzte Software (*mice* oder *rpart*). Wie in Abschnitt 2 angesprochen, unterscheiden sich die beiden Programmpakete unter anderem in Bezug auf die Behandlung fehlender Werte in den Prädiktoren. Deren Ausfallraten im Simulationsdatensatz sind eher gering. Unter

10 Der Mikrozensus-Namenszusatz „EF" steht für „Eingabefeld".

11 Seitens des Statistischen Bundesamtes wurde dieser Vergleich bisher nicht angestellt. Spies und Lange (2018) beschreiben erste Analysen zur Imputation der dichotomen Mutterschaftsvariable mit verschiedenen Machine Learning-Verfahren einschließlich CART, wenn auch ohne systematischen Methodenvergleich mit klassischen Methoden.

12 Die Originalsyntax zur PMM-Imputation wurde der Autorin dankenswerterweise vom Statistischen Bundesamt zur Verfügung gestellt, um den Methodenvergleich durchführen zu können.

den genannten Prädiktorvariablen sind nur sieben überhaupt von Antwortausfällen betroffen. Der maximale Ausfall liegt bei rund 2,6 % für die Einkommensvariable.

Dass *mice* Fälle mit unvollständigen Prädiktorvariablen von der Imputation ausschließt – es sei denn, sie werden zuvor selbst vervollständigt – ist nicht CART-spezifisch, sondern gilt auch für Imputationen mit *Predicitive Mean Matching*. Um dennoch Imputationswerte für alle Fälle mit fehlenden Werten in der Responsevariable zu erhalten, reagiert Destatis mit einer Umkodierung: Fälle mit fehlenden Werten in den Prädiktorvariablen werden entweder als eigene gültige Kategorie „ohne Angabe" kodiert oder zur gültigen Kategorie der „Sonstigen" gezählt. Dadurch sind diese Prädiktoren „vervollständigt" und alle Fälle können für den Imputationsprozess genutzt werden. Für die *mice*-basierten CART-Varianten wurde hier die gleiche Umkodierungsstrategie gewählt.

Da der zugrunde liegende, jedoch unbekannte, Wert eigentlich weder eine eigene Kategorie bildet noch in allen Fällen zu den „Sonstigen" gehören dürfte, ist die Strategie der Umkodierung nicht optimal, da so falsch zugeordnete Fälle in das Imputationsmodell einfließen könnten. Durch die Umkodierung stehen solche Fälle nun zwar für den Modell- bzw. für den Imputationsschritt zur Verfügung, allerdings um den Preis einer gewissen Modellverschlechterung – auch wenn sich der Anteil fehlender Werte, wie gesehen, in engen Grenzen hält.[13]

Im Gegensatz zu *mice* behält *rpart* Fälle mit fehlenden Prädiktorwerten bei. Für die Splitberechnungen zur Baumgenerierung im Modellschritt stehen zwar entsprechend weniger Fälle zur Verfügung, dafür ist das Modell nicht durch eventuelle Fehlkodierungen beeinträchtigt. Im Imputationsschritt können für fehlende Splitvariablen Surrogatsplits genutzt werden.[14] Für die *rpart*-basierten CART-Varianten konnte daher auf die Umkodierung der fehlenden Werte in den Prädiktorvariablen verzichtet werden.

Ein weiterer Unterschied zwischen beiden Softwarevarianten ist, dass in *mice* – unabhängig von der genutzten Imputationsmethode – eine Dichotomisierung von kategorialen Prädiktorvariablen zu Dummyvariablen erfolgt, kombiniert mit der Aussonderung redundanter Prädiktoren. Die Dichotomisierung ist (wie die Vermeidung von Redundanz bzw. Multikollinearität) für regressionsbasierte Imputationsmethoden notwendig, in Bezug auf CART werden dadurch jedoch die Splitoptionen je Entscheidungsebene unnötig eingeschränkt.[15]

Inwieweit sich die unterschiedlichen Herangehensweisen merklich auf die Imputationsergebnisse im gegebenen Anwendungsfall auswirken, konnte durch die Differenzierung der Vergleichsvarianten nach genutzter Software untersucht werden.

In den Simulationsdatensatz gingen nur diejenigen Frauen im Alter von 15 bis 75 Jahren ein, für die bekannt war, ob und ggf. wie viele Kinder sie geboren haben. Dies waren nach einer Bereinigung der Daten 253 086 Fälle. Mit diesen Fällen wurden nun 200 Simulationsdurchgänge durchgeführt. In jedem dieser 200 Durchgänge wurden die folgenden drei Schritte durchlaufen:

1. **Datenlöschung**: Bei rund 13,5 % der Fälle wurden deren Angaben zur Anzahl der Kinder gelöscht. Die reale Ausfallrate von rund 9 % wurde um das 1,5-fache überschritten, um noch kontrastreichere Ergebnisse aus der Untersuchung zu gewinnen, und auch da höhere Ausfallraten in Zukunft nicht ausgeschlossen sind. Um den Datenausfall realistisch abzubilden, wurden die Daten zufällig, aber in Abhängigkeit von Einflussvariablen gelöscht, die auch in der Realität eine Rolle spielen (wie bereits angesprochen, sind die Ausfallraten bei bestimmten Personengruppen größer als bei anderen).[16]
2. **Datenimputation:** Der unvollständige Datensatz wurde den fünf Vergleichsmethoden in identischer Form zugewiesen, sodass alle separat vor der identischen Aufgabenstellung standen, die fehlenden Werte zu ergänzen. Während für PMM, wie von Destatis praktiziert, in zwölf Altersklassen imputiert wurde, wurden für die CART-Varianten doppelt so große, also sechs statt zwölf Altersklassen als Imputationsklassen gewählt. Dadurch wurde die jeweilige Datenbasis informativer. Die Nutzung von Imputationsklassen wirkt im CART-Kontext wie eine „Vor-Partitionierung" der Daten und spart Rechenkapazität.
3. **Berechnung der Analysewerte:** Für den Simulationsdatensatz wurden die Verteilungen vor der Löschung und nach der Imputation anhand ausgewählter Untersuchungsgrößen miteinander verglichen und die Ergebnisse jedes Simulationslaufs dokumentiert. Als Analysegrößen dienten diverse Anteils- und Mittelwerte unterschiedlich tiefer Gliederung, die auch in den Destatis-Standardveröffentlichungen auftreten:

Analysegrößen der Simulationsstudie

- Anteile der Frauen mit 0 bis 9 oder mehr Kindern
- Anteile der Frauen mit 0 bis 5 oder mehr Kindern nach Altersklasse
- Anteile der Frauen mit 0 bis 3 oder mehr Kindern in der Altersklasse 45-49 nach Bundesland

13 Eine alternative Strategie wäre es, alle unvollständigen Variablen verkettet zu imputieren, das heißt, aufbauend auf zufälligen Startwerten, eine vorläufig vervollständigte Responsevariable in der nächsten Sequenz als Prädiktorvariable zu nutzen und diese Schrittfolge iterativ zu wiederholen, bis nach genügend Iterationen der Einfluss der Startwerte eliminiert ist. Da hier jedoch der Vergleich zwischen der tatsächlich praktizierten Destatis-PMM-Variante und der Nutzung eines alternativen CART-Modells im Vordergrund steht, wurde zwecks besserer Vergleichbarkeit auf eine Veränderung der Rahmenbedingungen verzichtet.

14 Alternativ sind natürlich auch mit *rpart* verkettete Imputationen möglich.

15 Ohne Dichotomisierung kann z. B. ein Split einer kategorialen Prädiktorvariable mit vier Ausprägungen so gestaltet sein, dass die Ausprägungen {1, 3} in den linken und die Ausprägungen {2, 4} in den rechten Tochterknoten weitergeleitet werden. Mit Dichotomisierung bleiben auf derselben Ebene nur Möglichkeiten der Art, dass ein Tochterknoten durch eine Ausprägung (dann „1"), der andere durch alle anderen (dann „0") definiert wird. Gegenüber Konkurrenzsplits auf derselben Ebene kann dies die suboptimale Variante sein.

16 Um den Einfluss dieser Variablen auf den Ausfall zu quantifizieren, wurde mit den ursprünglichen Daten ein Ausfallmodell geschätzt. Das Modell wurde dann genutzt, um die Ausfallwahrscheinlichkeiten der Fälle im Simulationsdatensatz zu berechnen. Auf Basis der jeweiligen Ausfallwahrscheinlichkeit entschied ein Zufallszug, ob ein Fall gelöscht wurde oder nicht.

- Anteile der Frauen mit 0 bis 3 oder mehr Kindern in der Altersklasse 45-49 nach Bildungsniveau
- Anteile der Frauen mit 0 bis 3 oder mehr Kindern in der Altersklasse 45-49 und in städtischem Gebiet nach Bildungsniveau
- Anteile der Frauen mit 0 bis 3 oder mehr Kindern in der Altersklasse 45-49 und in ländlichem Gebiet nach Bildungsniveau
- Durchschnittliche Kinderzahl der Mütter nach Altersklasse
- Durchschnittliche Kinderzahl der Mütter in der Altersklasse 45-49 nach Bundesland

Nach Abschluss der Simulationsdurchgänge wurden für jede Analysegröße die folgenden zusammenfassenden Qualitätsmaße berechnet:

Qualitätsmaße der Simulationsstudie

- die über die S=200 Simulationen gemittelte Abweichung zwischen dem Wert der Analysegröße nach der Imputation und dem Wert berechnet auf Basis der Originaldaten
- die Schwankung der Werte der interessierenden Analysegröße nach der Imputation, bemessen anhand der Standardabweichung über die S=200 Simulationen
- die mittlere *Data Utility (DU)* als Maßzahl für eine gesamte Tabelle

$$mittlere\ DU_{Tab} = \frac{1}{S}\sum_{s=1}^{S} DU_{Tab,S}$$

$$\text{mit } DU_{Tab,S} = \sum \left(Zellwert_{Tab\ imputiert\ ,S} - Zellwert_{Tab\ original\ ,S}\right)^2$$

Wünschenswert sind mittlere Abweichungen um Null zwischen den Ergebnissen auf Basis der vervollständigten Daten und den Ergebnissen auf Basis der Originaldaten. Zusätzlich interessiert, wie stark die Schätzergebnisse um ihren Mittelwert schwanken. Daher wurde auch die Standardabweichung der Schätzergebnisse betrachtet. Beide Maße beziehen sich auf Werte in Tabellenzellen. Als übersichtliche Maßzahl für eine gesamte Tabelle wurde in Anlehnung an Shlomo et al. (2010) die Maßzahl der mittleren Data Utility (DU) berechnet. Dazu wurden in jedem Simulationsdurchgang die quadrierten Abweichungen je Tabelle aufsummiert und die Ergebnisse je Tabelle über alle Simulationen gemittelt. Ein kleiner DU-Wert entspricht also einem besseren Ergebnis. Vergleichbar sind die Werte unterschiedlicher Methoden für jede Untersuchungstabelle.

5. Ergebnisse der Simulationsstudie

Den besten Gesamtüberblick bietet zunächst Tabelle 1, in der die mittleren Data Utility-Werte der verglichenen Imputationsmethoden dokumentiert sind. Für jede Untersuchungsgröße zeigt die rote Zellfärbung die schlechteste Performanz an, die grüne die insgesamt beste Performanz, die gelbe die beste Performanz unter den Varianten, die sich auf das PMM-Prädiktorvariablenset beschränken.

Die Ergebnisse sind insofern eindeutig, als dass im untersuchten Anwendungskontext die Performanz der CART-Varianten der von PMM überlegen ist. Offenbar kann die Komplexität im untersuchten Datensatz insgesamt besser durch eine CART-Modellierung abgebildet werden. Dies gilt bereits, wenn das zu PMM identische Prädiktorvariablenset genutzt wird. Die bigmodel-Varianten zeigen jedoch die insgesamt beste Performanz. Auch in Bezug auf die Schwankung der Schätzgrößen zeigt die PMM-Imputationsmethode im untersuchten Anwendungsfall eine häufig schlechtere Performanz als die CART-Verfahren.

Besonders große relative Performanzschwächen der PMM-Methode treten bei den Untersuchungsgrößen auf, die nach allen Altersklassen differenzieren. Werden die entsprechenden Ergebnisse im Detail betrachtet, so fällt auf, dass gegenüber den CART-Varianten besondere Probleme bestehen, die Untersuchungsgrößen in den Altersklassen ab 60 Jahren korrekt einzuschätzen. Abbildung d zeigt dies exemplarisch für die Anteile der Frauen mit 0 bis 5 oder mehr Kindern nach Altersklasse. Demnach bilden die CART-Varianten die Variablenbeziehungen vor allem in diesen Altersklassen besser ab als das PMM-Modell.

Weniger eindeutig gestaltet sich der Vergleich zwischen den CART-Varianten, wenn es um die Differen-

1 | Mittlerer Data Utility-Wert der verglichenen Imputationsmethoden

	PMM.*mice*	CART.*mice*	CART.*rpart*	CART.*mice. bigmodel*	CART.*rpart. bigmodel*
Anteile Frauen mit 0 bis 9 oder mehr Kindern in %	0,4611	0,0086	0,0120	0,0067	0,0077
Anteile Frauen mit 0 bis 5 oder mehr Kindern nach Altersklasse in %	54,2239	0,8629	0,8865	0,6720	0,6558
Anteile Frauen mit 0 bis 3 oder mehr Kindern in Altersklasse 45–49 nach Bundesland in %	59,6021	49,7463	43,1024	39,6923	28,4794
Anteile Frauen mit 0 bis 3 oder mehr Kindern in Altersklasse 45–49 nach Bildung in %	5,0125	1,1747	1,1278	1,0010	0,8125
Anteile Frauen mit 0 bis 3 oder mehr Kindern in Altersklasse 45–49 in städtischem Gebiet nach Bildung in %	7,5008	3,1629	3,1278	2,8074	2,2501
Anteile Frauen mit 0 bis 3 oder mehr Kindern in Altersklasse 45–49 in ländlichem Gebiet nach Bildung in %	8,0543	4,4761	4,8275	3,5860	3,3128
Durchschnittliche Kinderzahl der Mütter nach Altersklasse	0,0332	0,0021	0,0026	0,0018	0,0024
Durchschnittliche Kinderzahl der Mütter in Altersklasse 45–49 nach Bundesland	0,0092	0,0072	0,0076	0,0049	0,0048

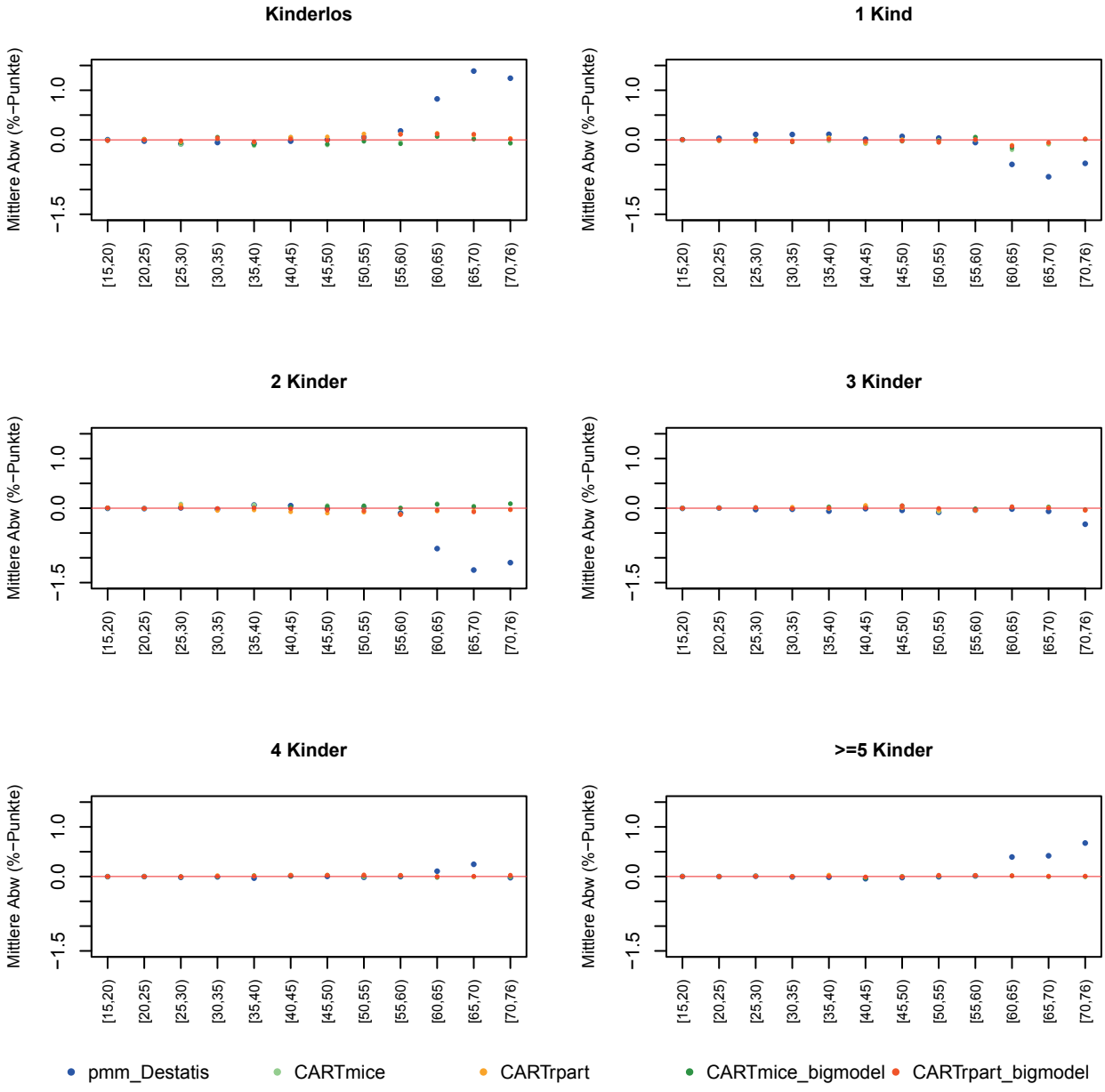
zierung nach der genutzten Software geht (vergleiche dazu noch einmal Tabelle 1). Vergleichbar sind die Varianten mit dem gleichen Imputationsmodell. Die Unterschiede hinsichtlich der DU sind jeweils nicht groß. Unter Hinzunahme der Zusatzprädiktoren im bigmodel entwickelt sich die Performanz aber etwas stärker zugunsten der CART.*rpart*-Variante.

Ein für weitergehende Analysen interessantes Zusatzfeature im *rpart*-Paket ist die Möglichkeit, sich die prozentuale *Variable Importance* ausgeben zu lassen: Damit wird der prozentuale Beitrag jeder Variable zur Gesamtreduktion der Heterogenität ausgewiesen. Wie Abbildung e (mit den über alle Simulationen gemittelten Werten) zeigt, leisteten die zusätzlichen Prädiktoren hier durchaus relevante Beiträge. Dies trifft insbesondere auf die Variable EF770 (Zahl der ledigen Kinder in der Familie/Lebensform; weniger in der Altersklasse ab 65 Jahren) und mit Abstrichen auch auf die Variablen EF401 und EF2009 (überwiegender Lebensunterhalt bzw. Migrationsstatus; stärker in den Altersklassen ab 55 Jahren) zu, während die Variable EF120 (Führungs- oder Aufsichtskraft) kaum zusätzlich zur Erklärungskraft des Modells beigetragen konnte.

6. Fazit

CART-basierte Methoden lassen sich für die modellgestützte Imputation fehlender Werte sowohl in kategorialen als auch in metrischen Responsevariablen nutzen. Sie können helfen, Fehlspezifikationen im Imputationsmodell durch die algorithmische Identifikation komplexer Datenstrukturen zu vermeiden. Der Algorithmus eignet sich insbesondere dafür, in großen Datensätzen mit einer Vielzahl von Prädik-

d | Mittlere Abweichung der Anteile der Frauen mit 0 bis 5 oder mehr Kindern nach Altersklasse

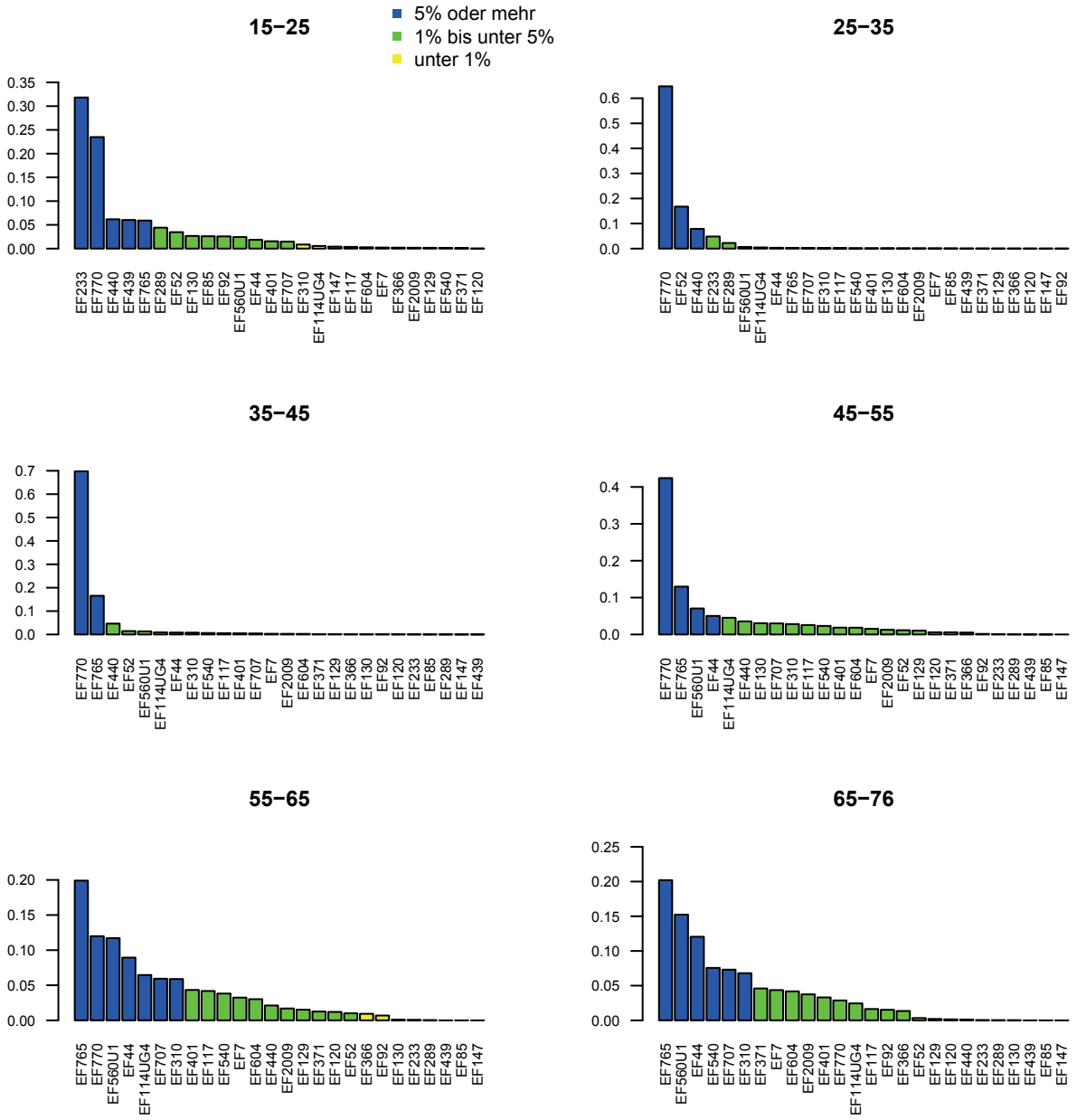


toren Interaktionseffekte auch höherer Ordnung und nichtlineare Datenzusammenhänge adaptiv zu ermitteln. Ein weiterer Vorteil ist, dass Multikollinearität für die Anwendung des CART-Algorithmus unproblematisch ist und sogar zum Vorteil genutzt werden kann. Wenn im Datensatz lineare und additive Strukturen vorherrschen, gelten korrekt spezifizierte parametrische Imputationsmodelle dem CART-Algorithmus allerdings als überlegen, da solche Strukturen durch die CART-Partitionierungen weniger gut abgebildet werden können. Methodenvergleiche in Simulationen mit den interessierenden Anwendungsdaten erlauben es, zusätzliche Erkenntnisse über die vorliegenden Datenstrukturen bzw. über die Eignung unterschiedlicher Imputationsmethoden zu gewinnen.

Illustriert wurde dies anhand einer Simulationsstudie zur Imputation fehlender Werte in der Variable „Anzahl geborener Kinder" mit Daten des Mikrozensus 2016. Verglichen wurde das vom Statistischen Bundesamt genutzte Predictive Mean Matching mit vier verschiedenen CART-Varianten, die sich hinsichtlich des jeweils genutzten Prädiktorvariablensets (Destatis-Auswahl versus erweiterte Auswahl) und in Bezug auf die genutzte Software (*mice* versus *rpart*) unterschieden. Differenzen zwischen den Softwarevarianten bestehen vor allem hinsichtlich der Behandlung fehlender Werte in den Prädiktorvariablen sowie in Bezug auf die Dichotomisierung kategorialer Prädiktoren zu Dummyvariablen.

Die Ergebnisse sind insofern eindeutig, als dass im untersuchten Anwendungskontext die Performanz

e | **Mean Variable Importance nach Altersklassen** *(rpart, bigmodel)*



beider CART-Varianten der von PMM überlegen ist. Offenbar kann die Komplexität im untersuchten Datensatz insgesamt besser durch eine CART-Modellierung abgebildet werden. Dies gilt bereits, wenn das identische Prädiktorvariablenset genutzt wird. Die *bigmodel*-Varianten zeigten jedoch die insgesamt beste Performanz. Wie weitere Analysen veranschaulichten, konnten insbesondere drei der in das erweiterte Prädiktorvariablenset aufgenommenen Variablen zur Erklärungskraft des Modells beitragen.

Weniger eindeutig gestaltet sich im gegebenen Anwendungsbeispiel der Vergleich zwischen den CART-Varianten, wenn es um die Differenzierung nach der genutzten Software geht. Die Ergebnisse geben jedoch Hinweise darauf, dass die Nutzung der *rpart*-Variante gegenüber der *mice*-unterstützten Imputation mit CART vorteilhaft sein kann, um die erweiterten Optionen im Umgang mit fehlenden Werten und mit kategorialen Variablen zahlreicher Ausprägungen in *rpart* nutzen zu können. Die Ergebnisse unterschieden sich zwischen beiden Softwareoptionen jedoch nicht gravierend. Aus Praktikabilitätserwägungen sollte auch beachtet werden, dass sich die Nutzung der *mice*-unterstützten CART-Imputation weniger programmierintensiv gestaltet, insbesondere wenn komplexere Imputationsvorhaben (verkettete Imputation, multiple Imputation) geplant sind. Hinsichtlich des Zeitaufwandes arbeiteten die CART.*rpart*-Varianten wiederum effizienter.[17]

Im Gesamtfazit bleibt festzuhalten, dass es auch für Imputationsvorhaben in der amtlichen Statistik lohnenswert erscheint, CART-basierte Modelle zu berücksichtigen und in vorbereitende statistische Methodenvergleiche mit einzubeziehen.

Birgit Pech ist Referentin im Referat *Mikrozensus, Sozialberichte* des Amtes für Statistik Berlin-Brandenburg.

Literatur

- Berk, R. A. (2008): Statistical Learning from a Regression Perspektive, New York: Springer.
- Breiman, L.; Friedman, J.; Stone, C. J. und Olshen, R. A. (1984): Classification and Regression Trees, Chapman and Hall/CRC.
- Burgette, L. F. und Reiter, J. P. (2010): Multiple Imputation for Missing Data via Sequential Regression Trees, in: Americal Journal of Epidemiology, 172: 1070–1076.
- Doove, L. L.; van Buuren, S. und Dusseldorp, E. (2014): Recursive partitioning for missing data imputation in the presence of interaction effects, in: Computational Statistics and Data Analysis, 72: 92–104.
- Drechsler, J. und Reiter, J. P. (2011): An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets, in: Computational Statistics and Data Analysis, 55: 3232–3243.
- Gaffert, P.; Meinfelder, F. und Bosch, V. (2016): Towards an MI-proper Predictive Mean Matching, Nürnberg/Bamberg 2016.
- Hastie, T.; Tibshirani, R. und Friedman, J. (2008): The Elements of Statistical Learning. Data Mining, Inference, and Prediction, Second Edition, New York: Springer.
- James, G.; Witten, D.; Hastie, T. und Tibshirani, R. (2013): An Introduction to Statistical Learning with Applications in R, New York: Springer.
- Little, R. J. A. (1988): Missing data adjustments in large surveys (with discussion), Journal of Business Economics and Statistics, 6, 287–301.
- Loh, W.-Y. (2014): Fifty Years of Classification and Regression Trees, in: International Statistical Review (2014), 82 (3): 329-348.
- Loh, W.-Y.; Eltinge, J.; Cho, M. J. und Li, Y. (2019): Classification and regression trees and forests for incomplete data from sample surveys. Statistica Sinica 29 (2019), 431–453.
- Meinfelder, F. (2009): Analysis of Incomplete Survey Data – Multiple Imputation via Bayesian Bootstrap Predictive Mean Matching. Disstertation, Bamberg.
- Reiter, J. P. (2005): Using CART to Generate Partially Synthetic Public Use Microdata, in: Journal of Official Statistics, 21(3): 441–462.
- Spies, L. und Lange, K. (2018): Implementation of artificial intelligence and machine learning methods within the Federal Statistical Office of Germany, Working Paper, Workshop on Statistical Data Editing, Neuchâtel 18–20 September 2018.
- Shlomo, N.; Tudor, C. und Groom, P. (2010): Data Swapping for Protectioning Census tables, in: Domingo-Ferrer, J. und Magkos, E. (Eds.): Privacy in Statistical Databases. UNESCO Chair in Data Privacy International Conference, PSD 2010, Corfu, Greece, September 2010, Proceedings, Berlin/Heidelberg/ New York: Springer.
- Therneau, T. und Atkinson, B. (2019): *rpart*: Recursive Partitioning and Regression Trees. R package version 4.1-15. https://cran.r-project.org/web/packages/rpart/index.html
- van Buuren, S. und Groothuis-Oudshoorn, K. (2019): *mice*: Multivariate Imputation by Chained Equations. R package version 3.6.0, https://cran.r-project.org/web/packages/mice/index.html

17 Die 200 Simulationen im Ursprungsmodell benötigten für die CART.*rpart*-Variante eine Rechenzeit von knapp zwei Stunden, während die CART.*mice*-Variante für dieselbe Aufgabenstellung fast das Dreifache erforderte.