Thurow, Maria

Article

Optionen zur Bemessung des Abstandes zweier Verteilungen in der Praxis

WISTA - Wirtschaft und Statistik

Provided in Cooperation with:
Statistisches Bundesamt (Destatis), Wiesbaden

Suggested Citation: Thurow, Maria (2022) : Optionen zur Bemessung des Abstandes zweier Verteilungen in der Praxis, WISTA - Wirtschaft und Statistik, ISSN 1619-2907, Statistisches Bundesamt (Destatis), Wiesbaden, Vol. 74, Iss. 2, pp. 19-29

This Version is available at:
https://hdl.handle.net/10419/253428

OPTIONEN ZUR BEMESSUNG DES ABSTANDES ZWEIER VERTEILUNGEN IN DER PRAXIS

Maria Thurow

↘ **Schlüsselwörter:** fehlende Werte – Imputationsmethoden – Imputationsgüte – Verteilungsähnlichkeit – Campus-Files

ZUSAMMENFASSUNG

Der Umgang mit unvollständigen Datensätzen stellt einen wichtigen Aspekt der Datenaufbereitung dar. Häufig ist nicht bekannt, für welche Auswertung ein aus dem Bereich der amtlichen Statistik stammender Datensatz in Wissenschaft und Forschung verwendet wird. Sinnvoll ist daher, die Verteilung der ursprünglichen Daten möglichst gut zu reproduzieren. Hierfür wäre eine möglichst geeignete Imputationsmethode zu wählen. Ziel der diesem Beitrag zugrunde liegenden Arbeit ist es, in einer Simulationsstudie verschiedene Kenngrößen hinsichtlich ihrer Eignung zu beurteilen und die Güten verschiedener Imputationsmethoden anhand von zwei Beispieldatensätzen aus der amtlichen Statistik miteinander zu vergleichen.

↘ ***Keywords:*** *missing values – imputation methods – imputation accuracy – distributional similarity – Campus files*

ABSTRACT

Dealing with incomplete data sets is an important aspect of data processing. It is often not known for what analysis a set of data from official statistics will later be used in science and research. Therefore, it is reasonable to reproduce the distribution of the original data as closely as possible. To this end, it is important to choose the most appropriate imputation method. The thesis described in this article uses a simulation study to assess the suitability of different quantities for comparing the goodness of fit of different imputation methods, based on two example data sets from official statistics.

Maria Thurow

studiert im Masterstudiengang Statistik an der Technischen Universität Dortmund und ist zudem als wissenschaftliche Hilfskraft an der Fakultät Statistik tätig. Im vorliegenden Artikel stellt sie ihre Bachelorarbeit mit dem Titel „Optionen zur Bemessung des Abstandes zweier Verteilungen in der Praxis“ vor, für die sie 2021 mit dem Gerhard-Fürst-Preis des Statistischen Bundesamtes ausgezeichnet wurde. Die Arbeit entstand an der TU Dortmund unter der Betreuung von Prof. Dr. Markus Pauly und Dr. Florian Dumpert.

1

Einleitung

Bei der Datenaufbereitung stellt der Umgang mit fehlenden Werten in einem Datensatz einen wichtigen Aspekt dar. Ein möglicher Ansatz ist das Entfernen unvollständiger Beobachtungen aus dem Datensatz (Complete Case Analysis). Eine Alternative zu diesem Verfahren ist die Imputation, bei der plausible Werte fehlende Werte ersetzen. Die Wahl einer geeigneten Imputationsmethode hängt dabei von der späteren Verwendung des Datensatzes ab. Ist diese nicht bekannt, sollte die zugrundeliegende Verteilung der Daten durch die Imputation möglichst gut reproduziert werden und viele spätere Anwendungen ermöglichen. Eine Möglichkeit, eine geeignete Kenngröße zu bestimmen, mit deren Hilfe die Abstände der (univariaten) Verteilungen der Variablen der ursprünglichen und der imputierten Daten bestimmt werden können, ist die Durchführung von Simulationen.

Eine solche Simulation wird in der diesem Beitrag zugrunde liegenden Arbeit anhand zweier Datensätze aus dem Bereich der amtlichen Statistik durchgeführt. Hierzu werden zuerst fehlende Werte in zwei Datensätzen simuliert. Anschließend werden die unvollständigen Datensätze mithilfe verschiedener Imputationsmethoden imputiert. Im Anschluss daran erfolgt ein Vergleich der (empirischen) Verteilungen der Variablen der ursprünglichen und der imputierten Daten. Hierzu werden verschiedene Kenngrößen verwendet, mithilfe derer die Abstände von Verteilungen bestimmt werden können. Basierend auf den beobachteten Werten für die Kenngrößen können diese hinsichtlich ihrer Eignung, die Güten verschiedener Imputationsmethoden zu beurteilen, verglichen werden.

Die für die Simulation verwendeten Datensätze werden in Kapitel 2 kurz vorgestellt. Anschließend folgt in Kapitel 3 eine Einführung der verwendeten Methoden. Dazu gehören die betrachteten Abstandsmaße für univariate Verteilungen und die Imputationsgüten, Mechanismen, unter denen fehlende Werte in Daten entstehen können, sowie die verwendeten Imputationsmethoden. Den Aufbau der Simulation beschreibt Kapitel 4, die Auswertung der Ergebnisse enthält Kapitel 5. Zusammengefasst werden die Ergebnisse der Auswertung in Kapitel 6.

Die durchgeführte Simulation, die Auswertung und das Erstellen der Grafiken erfolgten mit der Statistik-Software R (R Core Team, 2020).[1]

2

Verwendete Daten

Bei den für die Simulation verwendeten Datensätzen handelt es sich um gemäß dem Bundesstatistikgesetz absolut anonymisierte und für die Verwendung an Hochschulen angepasste Datensätze (sogenannte Campus-Files). Diese werden von den Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder zur Verfügung gestellt (Zwick, 2008). Für die in diesem Aufsatz beschriebene Arbeit wurden der Arbeitnehmerdatensatz der Verdienststrukturerhebung aus dem Jahr 2010 sowie der Datensatz der nach Fallpauschalen abgerechneten vollstationären Krankenhausfälle in Deutschland 2010 (DRG-Statistik) genutzt. Beide Datensätze enthalten fortlaufende Identifikationsnummern für die Beobachtungen, die jedoch aus den Datensätzen entfernt wurden, da sie bei der Imputation keine Relevanz haben. Zudem weisen beide Datensätze fehlende Werte in mehreren Variablen auf. Möglicherweise umfassen diese Variablen ebenfalls bei der Imputation relevante Informationen; aus diesem Grund werden die Datensätze vor der Simulation aufbereitet, sodass in den für die Simulation verwendeten Datensätze keine fehlenden Werte auftreten. So schließt der Datensatz der DRG-Statistik eine Variable mit dem Aufnahmegewicht der Patientinnen und Patienten ein. Weil dieses jedoch lediglich für Patientinnen und Patienten erfasst wird, die maximal ein Jahr alt sind, treten viele fehlende Werte auf. Im Zuge der Datenaufbereitung wird für Patientinnen und Patienten, die älter als ein Jahr sind, eine eigene Kategorie eingeführt und bei den entsprechenden Krankenhausfällen angegeben.

Der Datensatz der DRG-Statistik umfasst nach der Datenaufbereitung 46 Variablen und der Arbeitnehmerdatensatz der Verdienststrukturerhebung 28 Variablen. Eine ausführliche Beschreibung der Variablen ist in den Metadaten der Datensätze zu finden, welche auf der Webseite der Forschungsdatenzentren zur Verfügung stehen.

1 In diesem Beitrag werden nur die grundlegenden Ideen der Methoden beschrieben. Eine detaillierte Methodenbeschreibung ist bei Thurow und andere (2021) zu finden.

3

Verwendete Methoden

3.1 Kenngrößen zur Beurteilung der Güte bei der Imputation

Um möglichst viele spätere Auswertungen zu ermöglichen, sollte die ursprüngliche Verteilung der Daten durch die gewählte Imputationsmethode möglichst gut wiederhergestellt werden. Um den Abstand der (univariaten) Verteilungen der ursprünglichen und der imputierten Daten zu bestimmen, werden verschiedene Kenngrößen betrachtet. Für die kategorialen Variablen werden das X^2-Assoziationsmaß und das darauf basierende Cramérs V verwendet. Durch diese beiden Kenngrößen kann eine Aussage über den Zusammenhang zwischen den ursprünglichen und imputierten Daten getroffen werden.

Um die Verteilungen im Fall metrischer Variablen miteinander zu vergleichen, können die Realisierungen der Teststatistiken für Tests auf Verteilungsgleichheit verwendet werden. Im Folgenden werden hierzu die Realisierung der Teststatistik des Kolmogorow-Smirnow-Tests, des Cramer-von-Mises-Tests und des Anderson-Darling-Tests jeweils für zwei Stichproben verwendet. Die Realisierungen der Teststatistiken können auch für ordinale Variablen berechnet werden. Darüber hinaus werden die Kullback-Leibler-Divergenz und die quantil-basierte Mallow's-L^2-Distanz als Kenngrößen betrachtet. Mithilfe der Kullback-Leibler-Divergenz ist eine Aussage darüber möglich, wie gut die ursprüngliche Verteilung durch die Verteilung der imputierten Daten approximiert wird.

Neben den beschriebenen Kenngrößen werden zudem die normierte Wurzel der mittleren quadratischen Abweichung (NRMSE) für metrische Variablen sowie die Fehlklassifikationsrate (PFC) für kategoriale Variablen als klassische Imputationsgüten betrachtet.

3.2 Imputationstechniken in R

Es gibt verschiedene Verfahren, die für die Imputation fehlender Werte verwendet werden können. Für die Simulation wurden Imputationsverfahren ausgewählt, die bereits in der Statistik-Software R implementiert und weit verbreitet sind. Bei der Imputation kann zwischen einfachen Imputationsmethoden und der sogenannten multiplen Imputationsmethode unterschieden werden. Bei der einfachen Imputation wird für einen unvollständigen Datensatz ein plausibler vollständiger Datensatz erzeugt. Im Fall der multiplen Imputation werden für einen Datensatz mehrere (in dieser Arbeit $m = 5$) plausible Datensätze erzeugt. Im Folgenden werden lediglich die Grundideen der Imputationstechniken erläutert (Näheres siehe Thurow und andere [2021]). Als einfache Imputationsverfahren werden im Folgenden die Naive Imputation und die Imputation basierend auf einem Random Forest betrachtet.

Bei der Naiven Imputation werden die fehlenden Werte durch den Mittelwert (für metrische Variablen) beziehungsweise durch den Modus (für kategoriale Variablen) der beobachteten Werte der Variable ersetzt. In der R-Funktion missForest aus dem gleichnamigen Paket (Stekhoven/Bühlmann, 2012) wird der nichtparametrische Ansatz des Random Forest verwendet, um fehlende Werte zu imputieren. In dem R-Paket missRanger (Mayer, 2019) ist eine auf dieser Methode basierende Version der Random-Forest-Imputation implementiert, die eine geringere Laufzeit hat als missForest. Deshalb wird bei der in Kapitel 4 beschriebenen Simulation missRanger verwendet.

Neben den beiden beschriebenen einfachen Imputationsmethoden werden als multiple Imputationsmethoden eine Bootstrap-basierte Imputationsmethode (Amelia II; Honaker und andere, 2011) sowie drei Varianten der „Multiple Imputation by Chained Equations“ (MICE; van Buuren/Groothuis-Oudshoorn, 2011) verwendet.

In der in Amelia implementierten Methode wird der sogenannte EMB-Algorithmus verwendet. Hierbei wird der EM-Algorithmus (EM: expectation-maximization) auf mehreren Bootstrap-Stichproben der Daten angewendet. Bei dem MICE-Algorithmus werden die fehlenden Werte einer Variablen immer mithilfe der übrigen Variablen des Datensatzes imputiert. Dabei kann die Imputationsmethode zwischen den Variablen eines Datensatzes variieren. Zur Verfügung stehen bereits viele in dem R-Paket implementierte Funktionen, von denen in diesem Beitrag lediglich die Optionen „Predictive Mean Matching“ (pmm), „Random Forest“ (rf) sowie die Imputation mithilfe der Bayesschen Linearen Regression (norm) verwendet werden.

3.3 Simulationsmodelle für fehlende Werte

Nach Little und Rubin (2002) kann das Auftreten fehlender Werte nach unterschiedlichen Mechanismen erfolgen, die auch in der Praxis häufig auftreten. In der durchgeführten Simulation werden fehlende Werte nach dem zufälligen Fehlen und dem komplett zufälligen Fehlen simuliert. Treten fehlende Werte komplett zufällig auf, handelt es sich um den MCAR-Mechanismus (MCAR: missing completely at random). In diesem Fall ist das Fehlen der Werte also unabhängig von den beobachtbaren Werten eines Datensatzes. In der durchgeführten Simulation werden fehlende Werte unter dem MCAR-Mechanismus mithilfe der Funktion prodNA aus dem R-Paket missForest (Stekhoven/Bühlmann, 2012) simuliert.

Ist das Fehlen der Werte hingegen abhängig von den beobachteten (nicht fehlenden) Werten des Datensatzes, handelt es sich um den MAR-Mechanismus (MAR: missing at random). Um zu untersuchen, ob sich die Resultate für unterschiedliche Simulationsmodelle für fehlende Werte unterscheiden, werden bei Teilen der Simulation fehlende Werte mancher Variablen nach dem MAR-Mechanismus simuliert. Die Simulation der fehlenden Werte erfolgt dabei nach zuvor bestimmten Einflüssen anderer Variablen, welche für die hier verwendeten Daten in Kapitel 4 beschrieben werden.

4 Aufbau der Simulation

Für die Simulation werden die bereits beschriebenen Datensätze der DRG-Statistik und der Verdienststrukturerhebung aus dem Jahr 2010 verwendet. Um die Laufzeit zu verringern, wird bei der DRG-Statistik nicht der gesamte Datensatz verwendet, sondern bei jeder Iteration der Simulation ein neuer Teildatensatz von 10 000 Beobachtungen zufällig (ohne Zurücklegen) gezogen.

Bei der Simulation werden zuerst die fehlenden Werte in den jeweiligen (Teil-)Datensatz eingebaut. Im Anschluss daran erfolgt die Imputation mit allen in Abschnitt 3.2 beschriebenen Imputationsmethoden. Für jeden imputierten Datensatz werden die in Abschnitt 3.3 aufgeführten Kenngrößen berechnet. In beiden Datensätzen werden fehlende Werte nur in ausgewählten Variablen simuliert und für die Datensätze werden unterschiedliche Anteile fehlender Werte verwendet. Die fehlenden Werte bei der DRG-Statistik werden lediglich entsprechend dem MCAR-Mechanismus eingebaut und es werden die Fehlend-Raten von 1 %, 5 %, 10 % und 20 % verwendet. Bei der Verdienststrukturerhebung werden fehlende Werte sowohl unter dem MCAR- als auch dem MAR-Mechanismus simuliert. Hierzu werden drei zuvor bestimmte Einflüsse simuliert. Es handelt sich hierbei um mögliche Zusammenhänge, die keinen direkten Zusammenhang mit den tatsächlichen Einflüssen aus den Originaldaten der Forschungsdatenzentren haben:

- Mit steigendem Alter sinkt die Wahrscheinlichkeit für einen fehlenden Wert beim (normierten) Bruttojahresverdienst.
- Je geringer die (normierte) Wochenarbeitszeit ist, desto höher ist die Wahrscheinlichkeit, dass die Zulage für besondere Arbeitszeiten nicht angegeben ist.
- Die unterschiedlichen Ausbildungen (nach dem international angewendeten ISCED[2]-Schlüssel) haben unterschiedliche Wahrscheinlichkeiten für das Auftreten fehlender Werte bei der Leistungsgruppe bei der Vergütung nach freier Vereinbarung.

Bei den Arbeitnehmerdaten der Verdienststrukturerhebung werden 1 %, 5 % und 10 % fehlende Werte simuliert.

Jeder Simulationsansatz (bestehend aus einem Datensatz, einem Fehlend-Mechanismus und einer Fehlend-Rate) wird 100-mal hintereinander durchgeführt. Die Imputation mit dem Mice-Algorithmus wird dabei jeweils dreimal durchgeführt, wobei die Imputationsmethode für die einzelnen Variablen variiert. Für die metrischen Variablen variiert die Imputations-Option zwischen pmm, norm und rf (im Folgenden durch Mice.PMM, Mice.Norm und Mice.RF gekennzeichnet). Die kategorialen Variablen werden immer mithilfe eines Random Forest imputiert. Bei den multiplen Imputationen werden $m = 5$ imputierte Datensätze erzeugt. Für jeden dieser Datensätze werden die Abstandsmaße und Imputationsgüten berechnet und für die Auswertung das arithmetische Mittel der fünf Werte betrachtet.

2 ISCED: International Standard Classification of Education (Internationale Standardklassifikation im Bildungswesen).

5 Ergebnisse

Die in Abschnitt 3.1 beschriebenen Kenngrößen verfolgen verschiedene Ansätze, um die Güte der Imputationsmethoden zu beurteilen. Durch die Imputationsgüten NRMSE und PFC wird bestimmt, wie gut die Werte der zugrunde liegenden Stichprobe durch die Imputationsmethoden reproduziert werden können. Der Fokus liegt hierbei auf den tatsächlichen Werten. Bei den betrachteten Realisierungen der Teststatistiken sowie der Mallow's-L^2-Distanz und der Kullback-Leibler-Divergenz wird die Ähnlichkeit der Verteilungen der ursprünglichen und der imputierten Daten beurteilt.

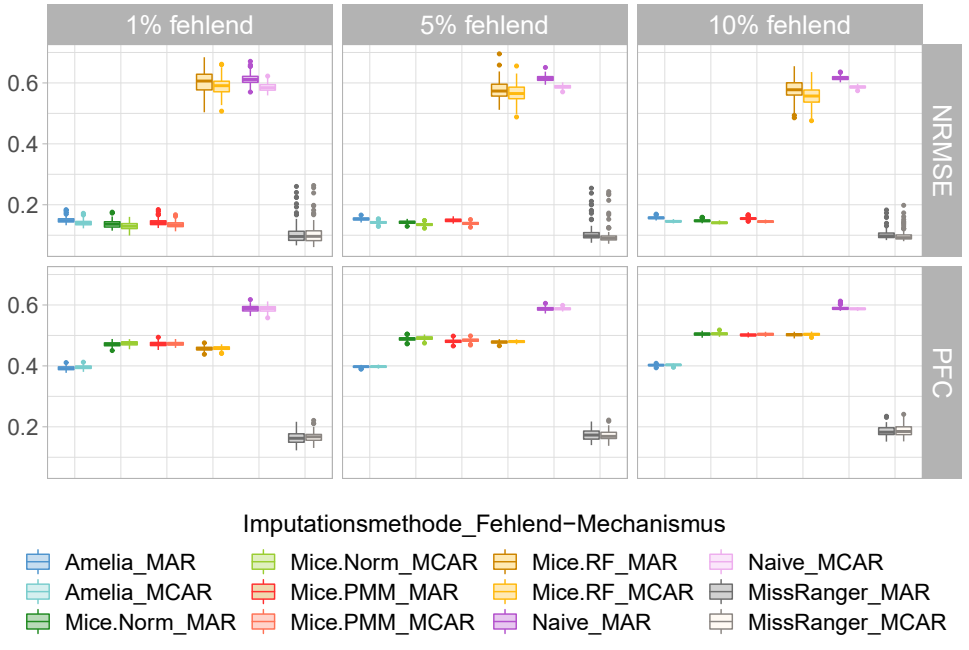
5.1 Imputationsgüten

In ↘ Grafik 1 sind Boxplots der beobachteten Werte des NRMSE und des PFC für den Datensatz der Verdienststrukturerhebung 2010 dargestellt. Sowohl für den NRMSE als auch den PFC deuten Werte nahe 0 nach Stekhoven und Bühlmann (2012) auf eine gute Imputation hin.

In Grafik 1 ist zu erkennen, dass sich die beobachteten Werte zwischen den beiden Simulationsmodellen für fehlende Werte nicht stark unterscheiden. Zudem verändern sich die Mediane der beobachteten Werte mit steigendem Anteil fehlender Werte nicht stark, die Variabilität sinkt jedoch geringfügig. Die niedrigsten Werte beider Kenngrößen können für die Imputation mit MissRanger beobachtet werden. Die höchsten Werte werden bei der Naiven Imputation erreicht. Für die drei Varianten des Mice-Algorithmus können für den PFC ähnliche Werte beobachtet werden. Grund hierfür ist vermutlich, dass bei allen bei der Simulation verwendeten Varianten des Mice-Algorithmus die kategorialen Variablen mit einem Random Forest imputiert werden. Beim NRMSE schneiden zwei der drei Varianten des Mice-Algorithmus (Mice.Norm und Mice.PMM) ähnlich gut und nicht viel schlechter als die Imputation mit MissRanger ab. Für den Mice-Algorithmus, bei dem auch die metrischen Varia-

Grafik 1

Boxplots der normierten Wurzel der mittleren quadratischen Abweichung (NRMSE) und der Fehlklassifikationsrate (PFC) für die verschiedenen Imputationsmethoden bei den Simulationen mit dem Arbeitnehmerdatensatz der Verdienststrukturerhebung 2010



2022 - 0098

blen mit einem Random Forest imputiert werden, können für den NRMSE ähnliche Werte beobachtet werden wie für die Naive Imputation. Die Imputation mit Amelia schneidet bei beiden Kenngrößen zwar schlechter ab als MissRanger, jedoch mindestens so gut wie die drei Varianten des Mice-Algorithmus.

Für die Simulation mit dem Datensatz der DRG-Statistik unterscheiden sich die zentralen Ergebnisse für den NRMSE und PFC nicht stark von den Ergebnissen bei der Verdienststrukturerhebung, weshalb sie an dieser Stelle nicht aufgeführt werden. Der einzige nennenswerte Unterschied ist, dass Amelia bei der Verdienststrukturerhebung bei dem PFC ähnlich schlecht abschneidet wie die Naive Imputation.

Werden lediglich die Imputationsgüten betrachtet, um die Imputationsmethoden zu vergleichen, lassen die Ergebnisse für beide Datensätze und sowohl für den NRMSE als auch den PFC vermuten, dass die Imputation mit MissRanger am besten geeignet ist.

Da die betrachteten Kenngrößen jedoch andere Ansätze für die Beurteilung der Imputationsmethoden verfolgen, sollten zusätzlich die Ergebnisse der Abstandsmaße für (univariate) Verteilungen betrachtet werden.

5.2 Abstandsmaße

Für die kategorialen Variablen beider Datensätze werden das X^2-Assoziationsmaß und das darauf basierende Cramérs V betrachtet. Da die Skala des X^2-Assoziationsmaßes von der Anzahl der Merkmalsausprägungen der Variablen abhängt, ist es für einen Vergleich der Imputationsmethoden sinnvoller, das auf X^2 basierende Cramérs V zu betrachten: Dieses nimmt Werte im Intervall [0,1] an, wobei Werte nahe 1 auf abhängige Verteilungen hindeuten. In der durchgeführten Simulation können für Cramérs V ähnliche Ergebnisse beobachtet werden wie für den NRSME und den PFC. Bei einem Vergleich der Imputationsmethoden sollte beachtet werden, dass das X^2-Assoziationsmaß und damit auch Cramérs V keine Kenngrößen für den Abstand von Verteilungen sind, sondern dazu dienen, die Abhängigkeit von Verteilungen zu beurteilen.

Um die Abstände der Verteilungen metrischer und ordinaler Variablen zu bemessen, werden die Realisierungen der Teststatistiken des Cramer-von-Mises-Tests, des Anderson-Darling-Tests und des Kolmogorow-Smirnow-Tests betrachtet. Darüber hinaus wird der Abstand der Verteilungen für metrische Variablen durch die Mallow's-L^2-Distanz und die Kullback-Leibler-Divergenz beurteilt.

Die Ergebnisse der betrachteten Kenngrößen unterscheiden sich nicht stark voneinander. Bei einigen Kenngrößen sind jedoch im Rahmen der Simulation Probleme und Nachteile aufgefallen, die dazu führen, dass sie nur bedingt für eine Beurteilung der Imputationsgüte geeignet sind:

Ein Nachteil der Kullback-Leibler-Divergenz ist, dass der hier verwendete Schätzer nach der Definition von Cover und Thomas (2006) in einigen Fällen den Wert ∞ annehmen kann. Dies ist in der Simulation der Fall, wodurch ein Vergleich der Ergebnisse schwierig ist. Die Mallow's-L^2-Distanz basiert auf der Quantil-Funktion. Dadurch hängt die Skala dieser Kenngröße von der Skala der jeweils betrachteten Variablen ab. Hierdurch ist es schwierig, für unterschiedliche Variablen zu beurteilen, wie gut die unterschiedlichen Imputationsverfahren abschneiden. Für die Realisierung der Teststatistik des Anderson-Darling-Tests wird in der hier beschriebenen Simulation ein rangbasierter Schätzer (Pettit, 1976) verwendet. Dessen Werte steigen bei Auftreten von Bindungen, also von doppelt auftretenden Werten in einer Stichprobe, an. Dies kann ungewollt zu schlechteren Ergebnissen bei der Auswertung führen, wenn einige Variablen zum Beispiel gerundet auftreten. Des Weiteren können hierdurch Imputationsverfahren, bei denen fehlende Werte durch beobachtete Werte imputiert werden (in diesem Beitrag zum Beispiel Mice.PMM), fälschlicherweise als schlechter beurteilt werden. Aus diesen Gründen werden die Ergebnisse für diese Kenngrößen nicht genauer beschrieben.

Für die beiden verbleibenden Kenngrößen, die Realisierungen der Teststatistiken des Cramer-von-Mises- und des Kolmogorow-Smirnow-Tests, sind keine großen Unterschiede bei den Ergebnissen zu erkennen. Die beobachteten Werte für die Kolmogorow-Smirnow-Teststatistik sind etwas stabiler als für die Teststatistik des Cramer-von-Mises-Tests. Letztere ist vermutlich etwas empfindlicher gegenüber leichten Veränderungen in den Daten. Da auch bei anderen Kenngrößen stabile Werte zu beobachten sind, werden im Folgenden lediglich die Ergebnisse für die Realisierung der Teststatistik des Kolmogorow-Smirnow-Tests ausführlich beschrieben.

Die Realisierung der Teststatistik des Kolmogorow-Smirnow-Tests basiert, anders als der NRMSE und PFC, auf den (empirischen) Verteilungsfunktionen der Variablen vor und nach der Imputation. Niedrige Werte weisen hierbei auf ähnliche Verteilungen hin.

In ↘ Grafik 2 sind Boxplots der beobachteten Werte für die Realisierung der Teststatistik des Kolmogorow-Smirnow-Tests für die ordinalen Variablen und die eine metrische Variable des Datensatzes der DRG-Statistik aufgeführt. Es ist zu sehen, dass die beobachteten Werte mit steigendem Anteil fehlender Werte ebenfalls ansteigen und dass für die meisten Imputationsmethoden auch die Variabilität der Werte zunimmt.

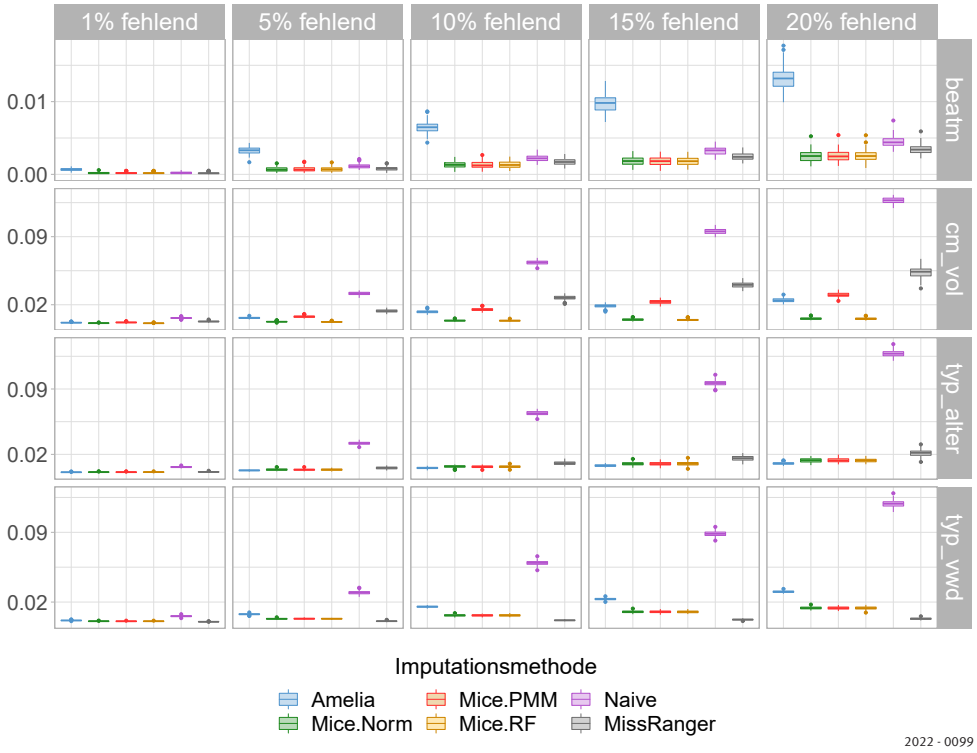
Die einzige metrische Variable des Datensatzes ist das Case-Mix-Erlösvolumen (cm_vol). Es fällt auf, dass es lediglich bei dieser Variablen zu auffälligen Unterschieden zwischen den drei Varianten des Mice-Algorithmus kommt. Der Unterschied liegt darin, dass die beobachteten Werte für Mice.PMM mit steigendem Anteil fehlender Werte stärker ansteigen als für die anderen Verfahren.

MissRanger schneidet für die verschiedenen Variablen unterschiedlich gut ab. Während diese Imputationsmethode beim Typ der Verweildauer einer Patientin/eines Patienten (typ_vwd) am besten abschneidet, schneidet sie beim gruppierten Alter (typ_alter) schlechter ab als die betrachteten Multiplen Imputationsmethoden (Amelia und Mice). Für die Naive Imputation können bei drei der vier Variablen mit Abstand die höchsten Werte beobachtet werden. Bei der Beatmungszeit (beatm) schneidet sie ähnlich gut ab wie die Mice-Verfahren und MissRanger. Bei dieser Variablen schneidet Amelia deutlich schlechter ab als die anderen Imputationsmethoden. Bei den anderen Variablen ist dieser Algorithmus jedoch ähnlich gut wie die anderen Verfahren.

↘ Grafik 3 enthält die Boxplots der beobachteten Realisierungen der Teststatistik des Kolmogorow-Smirnow-Tests bei den Simulationen mit den Arbeitnehmerdaten der Verdienststrukturerhebung. Ebenso wie bei der DRG-Statistik steigen die beobachteten Werte mit steigendem Anteil fehlender Werte an. Zudem ist zu erkennen, dass die beobachteten Werte für alle Fehlend-Raten

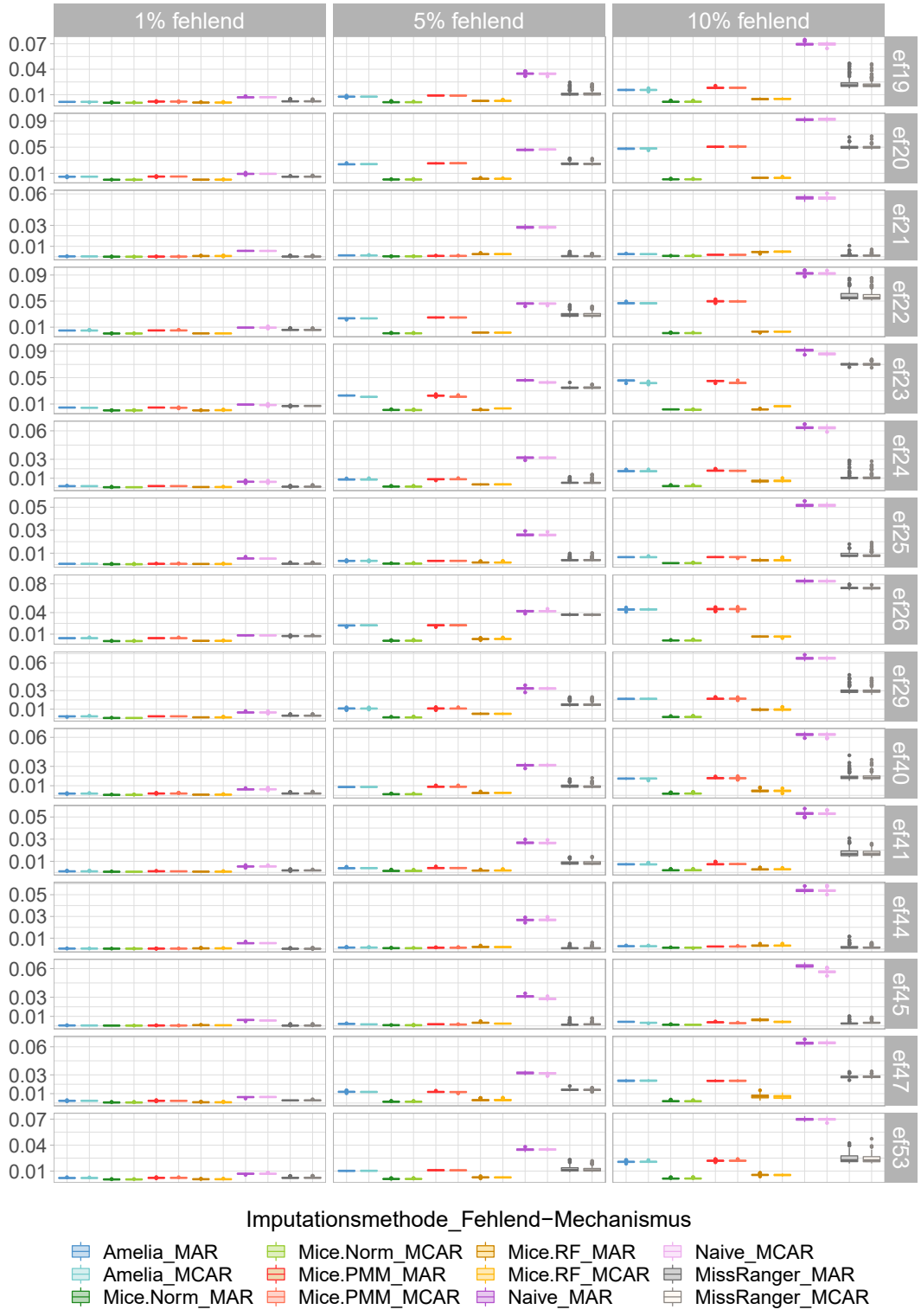
Grafik 2

Boxplots der beobachteten Werte für die Realisierung der Teststatistik des Kolmogorow-Smirnow-Tests für die ordinalen und die metrische Variable/-n bei der Simulation mit dem Datensatz der DRG-Statistik



Grafik 3

Boxplots der beobachteten Werte für die Realisierung der Teststatistik des Kolmogorow-Smirnow-Tests für die metrischen Variablen bei der Simulation mit dem Arbeitnehmerdatensatz der Verdienststrukturerhebung 2010



2022 - 0100

nicht stark variieren und keine großen Unterschiede zwischen den Fehlend-Mechanismen zu erkennen sind. Es ist außerdem zu sehen, dass bei der Imputation mit der Naiven Imputation immer die höchsten Werte für die Realisierung der Teststatistik auftreten. Die Werte bei der Imputation mit Amelia und Mice.PMM verhalten sich für die verschiedenen Variablen und Fehlend-Raten sehr ähnlich. Für fast alle Imputationsmethoden (außer MissRanger) sind keine großen Unterschiede der beobachteten Werte zwischen den Variablen zu erkennen. Für die Imputation mit Mice.Norm und Mice.RF können für alle Variablen sehr niedrige Werte beobachtet werden. MissRanger schneidet bei den unterschiedlichen Variablen unterschiedlich gut ab und es können bei keiner Variable deutlich niedrigere Werte beobachtet werden als für die Mice-Algorithmen.

Auffallend ist, dass die fünf Variablen ef21 (Bruttomonatsverdienst), ef24 (Lohnsteuer), ef25 (Sozialversicherungsbeiträge), ef44 (Nettomonatsverdienst) und ef45 (normierter Bruttojahresverdienst), bei denen sich die beobachteten Werte bei der Imputation mit MissRanger kaum von 0 unterscheiden, in direktem Zusammenhang zueinander stehen. Die fünf beschriebenen Variablen stehen alle in direktem Zusammenhang zum Einkommen. Entsprechend ist zu vermuten, dass fehlende Werte in einer Variable relativ zuverlässig durch die beobachteten Werte in anderen Variablen imputiert werden können. Dies ist vermutlich auch der Grund, weshalb für die anderen Imputationsmethoden (abgesehen von der Naiven Imputation) bei diesen Variablen auch niedrige Werte beobachtet werden können. Zudem fällt auf, dass sich die beobachteten Werte bei Mice.RF und MissRanger teilweise stark voneinander unterscheiden. Da die beiden Imputationsverfahren weitestgehend auf dem gleichen Algorithmus beruhen, ist dies erstaunlich. Bei einem Vergleich der drei Varianten des Mice-Algorithmus ist zu erkennen, dass die beobachteten Werte bei Mice.PMM für die meisten Variablen größer sind als bei den anderen beiden Varianten.

Die Ergebnisse der betrachteten Abstandsmaße unterscheiden sich teilweise von jenen der klassischen Imputationsgüten (NRMSE und PFC). Dies liegt vermutlich daran, dass durch die Imputationsgüten beurteilt werden kann, wie gut die tatsächlichen Werte reproduziert werden. Dagegen kann durch die betrachteten Abstandsmaße beurteilt werden, wie gut die Verteilung der ursprünglichen Daten durch die Imputationsmethoden reproduziert werden kann. Aufgrund der unterschiedlichen Ergebnisse sollten mehrere Kenngrößen für den Vergleich der Imputationsmethoden verwendet werden.

Neben den Distanzmaßen können anhand der Ergebnisse der durchgeführten Simulation auch die Imputationsmethoden verglichen werden. Die geringsten Werte bei der Simulation mit der DRG-Statistik sind bei der Teststatistik des Kolmogorow-Smirnow-Tests und Cramérs V für Mice.Norm und Mice.RF zu beobachten. Werden zusätzlich die Werte des NRMSE und PFC für eine Beurteilung verwendet, erfolgt unter Verwendung des Mice.RF-Algorithmus eine etwas bessere Imputation. Für den Arbeitnehmerdatensatz der Verdienststrukturerhebung schneiden bei den beiden Abstandsmaßen ebenfalls Mice.Norm und Mice.RF am besten ab. Beim NRMSE sind die beobachteten Werte für Mice.RF hier jedoch deutlich höher als für Mice.Norm. Bei dem PFC unterscheiden sich die Verfahren nicht stark voneinander. Bei einer zusätzlichen Analyse der Laufzeit, auf die in diesem Beitrag nicht explizit eingegangen wird, ist die Laufzeit von Mice.Norm niedriger als die von Mice.RF. Deshalb ist basierend auf den Ergebnissen der Simulation für die Imputation der beiden Datensätze insgesamt Mice.Norm Mice.RF vorzuziehen.

Dieses Ergebnis muss jedoch nicht für die Originaldatensätze der DRG-Statistik und der Verdienststrukturerhebung zutreffen, da die für diesen Beitrag verwendeten Datensätze in stark anonymisierter Form vorliegen und sich die Abstandsmaße und Imputationsmethoden bei den vollständigen Daten anders verhalten könnten.

6

Fazit

Ziel der in diesem Beitrag beschriebenen Arbeit ist, die Untersuchung verschiedener Abstandsmaße für univariate Verteilungen hinsichtlich ihrer Eignung sowie die Güten von Imputationsmethoden bei zwei ausgewählten Datensätzen der Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder zu beurteilen.

Für eine Beurteilung der Imputationsgüte bei kategorialen Variablen hinsichtlich der ursprünglichen Verteilung ist das X^2-Assoziationsmaß nicht geeignet. Grund dafür ist, dass durch dieses das Maß der Übereinstimmung

der imputierten und der originalen Daten, nicht aber die Ähnlichkeit der Verteilungen beurteilt werden kann.

Um die Imputationsgüte für ordinale und metrische Variablen zu bewerten, eignet sich bei den vorliegenden Datensätzen am besten die Realisierung der Teststatistik des Kolmogorow-Smirnow-Tests. Diese ist gut interpretierbar und die beobachteten Werte sind weitestgehend stabil, wobei die Unterschiede zur Realisierung der Teststatistik des Cramer-von-Mises-Tests minimal sind.

Es sollte beachtet werden, dass die Ergebnisse für andere Datensätze anders aussehen können. So enthält der Datensatz der DRG-Statistik nur eine metrische Variable und der Arbeitnehmerdatensatz der Verdienststrukturerhebung keine ordinalen Variablen. Außerdem sind die verwendeten Datensätze stark anonymisiert und die Beobachtungen einiger Variablen sind gerundet. Für die vollständigen (nicht anonymisierten) Daten könnten die Ergebnisse gegebenenfalls anders ausfallen.

Des Weiteren ist die Auswertung dieses Beitrags rein deskriptiv. Eine weiterführende Analyse in Thurow und andere (2021) betrachtet zusätzlich die p-Werte des Kolmogorow-Smirnow-Tests und weitere Kenngrößen, um die Anpassungsgüte bei der Imputation zu beurteilen. Für die kategorialen Variablen sollten außerdem andere Kenngrößen zur Bemessung des Abstandes der Verteilungen verwendet werden.

LITERATURVERZEICHNIS

Forschungsdatenzentren (FDZ) der Statistischen Ämter des Bundes und der Länder. DOI: 10.21242/23141.2010.00.00.5.1.0 und 10.21242/62111.2010.00.00.5.1.0, eigene Berechnungen.

Cover, Thomas M./Thomas, Joy A. *Elements of Information Theory*. 2. Auflage. New York 2006.

Honaker, James/King, Gary/Blackwell, Matthew. *Amelia II: A Program for Missing Data*. In: Journal of Statistical Software. Jahrgang 45. Ausgabe 7/2011, Seite 1 ff.

Little, Roderick J./Rubin, Donald B. *Statistical Analysis with Missing Data*. 2. Auflage. Chichester 2002.

Mayer, Michael. *missRanger: Fast Imputation of Missing Values*. R package version 2.1.0. 2019.

Pettit, Anthony. *A two-sample Anderson-Darling rank statistic*. In: Biometrika. Jahrgang 63. Ausgabe 1/1976, Seite 161 ff.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Wien 2020.

Stekhoven, Daniel J./Bühlmann, Peter. *MissForest – non-parametric missing value imputation for mixed-type data*. In: Bioinformatics. Jahrgang 28. Ausgabe 1/2012, Seite 112 ff.

Thurow, Maria/Dumpert, Florian/Ramosaj, Burim/Pauly, Markus. *Imputing Missings in Official Statistics for General Tasks – Our Vote for Distributional Accuracy*. In: Statistical Journal of the IAOS. Jahrgang 37. Ausgabe 4/2021, Seite 1379 ff.

van Buuren, Stef. *Flexible Imputation of Missing Data*. 2. Auflage. Boca Raton 2018.

van Buuren, Stef/Groothuis-Oudshoorn, Karin. *mice: Multivariate Imputation by Chained Equations in R*. In: Journal of Statistical Software. Jahrgang 45. Ausgabe 3/2011, Seite 1 ff.

Zwick, Markus. *CAMPUS-Files - Kostenfreie Public Use Files für die Lehre*. In: Wirtschafts- und Sozialstatistisches Archiv. Ausgabe 2/2008, Seite 175 ff.

Herausgeber
Statistisches Bundesamt (Destatis), Wiesbaden

Schriftleitung
Dr. Daniel Vorgrimler
Redaktion: Ellen Römer

Ihr Kontakt zu uns
www.destatis.de/kontakt

Erscheinungsfolge
zweimonatlich, erschienen im April 2022
Ältere Ausgaben finden Sie unter www.destatis.de sowie in der Statistischen Bibliothek.

Artikelnummer: 1010200-22002-4, ISSN 1619-2907