

Preising, Marcel; Lange, Kerstin; Dumpert, Florian

## Article

# Imputation zur maschinellen Behandlung fehlender und unplausibler Werte in der amtlichen Statistik

WISTA - Wirtschaft und Statistik

## Provided in Cooperation with:

Statistisches Bundesamt (Destatis), Wiesbaden

*Suggested Citation:* Preising, Marcel; Lange, Kerstin; Dumpert, Florian (2021) : Imputation zur maschinellen Behandlung fehlender und unplausibler Werte in der amtlichen Statistik, WISTA - Wirtschaft und Statistik, ISSN 1619-2907, Statistisches Bundesamt (Destatis), Wiesbaden, Vol. 73, Iss. 5, pp. 40-52

This Version is available at:

<https://hdl.handle.net/10419/244708>

### Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

### Terms of use:

*Documents in EconStor may be saved and copied for your personal and scholarly purposes.*

*You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.*

*If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.*

#### Dr. Marcel Preising

ist Statistiker und promovierte 2017 an der Universität Bamberg. Als Referent im Referat „Künstliche Intelligenz, Big Data“ des Statistischen Bundesamtes entwickelt er Verfahren zur Imputation und automatisierten Plausibilisierung – unter anderem aus dem Bereich des maschinellen Lernens. Zudem berät er bei deren Implementierung in die Statistikproduktion.

#### Kerstin Lange

hat Statistik studiert und ist als Referentin im Referat „Künstliche Intelligenz, Big Data“ des Statistischen Bundesamtes tätig. Ihr Arbeitsschwerpunkt liegt auf Imputationsverfahren.

#### Dr. Florian Dumpert

ist als Referent mit Verfahren des maschinellen Lernens und der Imputation im Referat „Künstliche Intelligenz, Big Data“ des Statistischen Bundesamtes befasst. Der Diplom-Mathematiker beschäftigt sich dabei unter anderem mit den methodischen Fragestellungen beim Einsatz dieser Verfahren.

# IMPUTATION ZUR MASCHINELLEN BEHANDLUNG FEHLENDER UND UNPLAUSIBLER WERTE IN DER AMTLICHEN STATISTIK

Theoretische Grundlagen und praktische Umsetzung

Marcel Preising, Kerstin Lange, Florian Dumpert

📌 **Schlüsselwörter:** Antwortausfall – fehlende Werte – Datenaufbereitung – Imputation – Plausibilisierung

## ZUSAMMENFASSUNG

Fehlende Werte in den Erhebungsdaten stellen für die amtliche Statistik eine Herausforderung dar. Ein falscher Umgang mit ihnen kann die Ergebnisse verzerren. Zudem machen große Datenmengen und hohe Anforderungen an die Aktualität der Daten zunehmend automatisierte Plausibilisierungs- und Imputationsverfahren notwendig. Um eine hohe Qualität der Ergebnisse erreichen zu können ist es daher wichtig, sich mit diesen Themen auseinanderzusetzen. Der Aufsatz beschreibt, warum fehlende Werte ein Problem darstellen können und zeigt einige Behandlungsoptionen auf. Er stellt verschiedene Klassen von Imputationsverfahren und Möglichkeiten zur Evaluation der Imputationsergebnisse vor und ergänzt sie mit Beispielen aus der amtlichen Statistik.

📌 **Keywords:** non-response – missing values – data processing – imputation – data editing

## ABSTRACT

*Missing values in survey data are a challenge for official statistics. Incorrect handling of them can distort the results. Furthermore, large amounts of data and high requirements on the timeliness of the results are increasing the need for automated data editing and imputation procedures. Therefore, it is important to deal with these issues in order to be able to achieve high quality results. The article describes why missing values can be a problem and discusses a number of options for addressing the issue. Different classes of imputation procedures and possibilities for evaluating the imputation results are presented and complemented with examples from official statistics.*

## 1

### Einleitung: Das Problem fehlender Werte

Die Statistischen Ämter des Bundes und der Länder stellen an ihre Statistikproduktion hohe Ansprüche: Nur wenn strenge Qualitätsanforderungen bei der Erstellung einer Statistik erfüllt wurden, erhalten statistische Ergebnisse das Prädikat „amtlich“. <sup>1</sup> Der typische Verlauf hin zur Veröffentlichung gemäß diesen Standards kann dem Geschäftsprozessmodell Amtliche Statistik (GMAS) entnommen werden (Gehle/Lüüs, 2017; Statistische Ämter des Bundes und der Länder, 2021).

Die fünfte Phase des GMAS umfasst die Aufbereitung der zu verarbeitenden Daten. Unabhängig davon, welcher Quelle die Daten entstammen, wird in dieser Phase sichergestellt, dass veröffentlichte Ergebnisse keinen unbereinigten Fehlern innerhalb der Daten unterliegen. Zentrale Fehlerquellen sind dabei unplausible oder fehlende Werte, die zum Beispiel in Antwortverweigerungen begründet liegen. <sup>2</sup>

Um diesen unplausiblen oder fehlenden Werten zu begegnen, existieren zahlreiche Behandlungsansätze. Einer hiervon ist die Imputation, bei der ein (fehlender) Wert unter (möglichst effizienter) Ausnutzung der sich (zumeist) im Datenbestand befindlichen Informationen durch einen anderen Wert ersetzt wird (Rubin, 1978; Rubin, 1987; Little/Rubin, 2002; Enders, 2010; van Buuren, 2018; Carpenter, 2013; de Waal und andere,

2011). Dieses Vorgehen weist gegenüber anderen Verfahren (siehe Kapitel 3) Vorteile durch die effizientere Nutzung vorliegender Informationen auf. Auch bietet es die Möglichkeit, nach seiner Durchführung auf eine vervollständigte Datentabelle, die auf der Nutzung aller zur Verfügung stehenden Informationen basiert, zugreifen zu können. Das erleichtert unter anderem die Berechnung von Kennzahlen mithilfe von Standardsoftware.

Neben den methodischen Vorteilen spricht noch ein weiteres Argument für eine Imputation innerhalb der Statistikproduktion: Sie kann als maschinelles Verfahren – und damit im Gegensatz zur manuellen Korrektur – auch große Datenmengen verarbeiten. In Zeiten steigender Datenumfänge und mit dem Ziel der schnelleren Veröffentlichung kommt der Imputation dabei insbesondere in der amtlichen Statistik eine immer größer werdende Bedeutung zu.

Entsprechend fand sie bereits in zahlreichen amtlichen Statistikproduktionen Anwendung, die teilweise auch schon länger zurückliegen (Beispiel: Esser und andere, 1989), und/oder von besonders hoher Relevanz waren (Beispiel: Grunwald/Krause, 2014). Imputation ist – nachdem die Daten zuvor als unplausibel oder fehlend identifiziert wurden <sup>3</sup> – fest vorgesehen innerhalb der GMAS-Phase 5.4 „Daten plausibilisieren und imputieren“.

Dieser Übersichts- und Grundlagenartikel bietet eine Einführung in die Imputation als maschinelles Verfahren zur Behandlung unplausibler und fehlender Werte. Er geht der Frage nach, wann und weshalb unplausible beziehungsweise fehlende Werte überhaupt ein Problem darstellen. Ebenfalls zeigt er auf, unter welchen Bedingungen ein Imputationsverfahren – auch mit Blick auf die einhergehende Unsicherheit, die sich bei der Berechnung von Standardfehlern innerhalb der Ergebnisausweisung widerspiegeln sollte – dieses beheben kann. Zudem wird erörtert, wie ein Imputationsverfahren vor, aber auch innerhalb einer laufenden Statistikproduktion zu evaluieren und zu dokumentieren ist.

1 Ergänzend dazu veröffentlicht das Statistische Bundesamt seit einiger Zeit in der Rubrik „EXDAT – Experimentelle Daten“ regelmäßig neue, innovative Projektergebnisse. Sie entstehen auf der Grundlage neuer Datenquellen und Methoden. Im Reifegrad und in manchen Aspekten der Qualität unterscheiden sie sich von amtlichen Statistiken, insbesondere in Bezug auf Harmonisierung, Erfassungsbereich und Methodik. Dennoch sind es Ergebnisse der Statistischen Ämter des Bundes und der Länder, die interessante, neue Perspektiven auf verschiedene Themenfelder der Statistik bieten.

2 Auch das Vorliegen von zwar erhobenen, jedoch unplausiblen Daten kann dadurch behandelt werden, dass die in die jeweilige Plausibilitätsregelverletzung involvierten Merkmale (gegebenenfalls partiell und regelbasiert) zunächst auf fehlend gesetzt werden, um sie anschließend zu imputieren. Der separate Vorgang, innerhalb dessen die Plausibilitätsregelverletzung festgestellt wird, ist jedoch nicht Gegenstand dieses Aufsatzes.

Als weiterer Grund für fehlende Werte kommen zum Beispiel Datenübermittlungsfehler infrage. Mitunter kann es aber auch geplant sein, dass gewisse Merkmale bei einem Teil der statistischen Einheiten nicht erhoben werden. Hierbei handelt es sich um sogenannte Missing-by-Design-Ausfallmechanismen.

3 Siehe GMAS-Phase 5.3 „Daten prüfen und validieren“. Es gibt jedoch auch Algorithmen, die eine gleichzeitige Fehlererkennung und Imputation vornehmen, wie die Software CANadian Census Editing and Imputation System (CANCEIS).

2

## Ausfallmechanismen

An dieser Stelle erfolgt zunächst eine begriffliche Einordnung der Form des Antwortausfalls: Fällt eine statistische Einheit (zum Beispiel ein Betrieb, ein Unternehmen, ein Haushalt, eine Person oder eine andere Einheit) samt allen zu erhebenden Merkmalen aus, so spricht man von Unit Nonresponse.<sup>14</sup> Fehlen hingegen lediglich einzelne Merkmale einer statistischen Einheit, handelt es sich begrifflich um den sogenannten Item Nonresponse, dem sich dieser Artikel im weiteren Verlauf widmet.

Der sogenannte Ausfallmechanismus ist nun von entscheidender Bedeutung für die Notwendigkeit, aber auch für den Erfolg eines Imputationsverfahrens (Rubin, 1978; Rubin, 1987; Little/Rubin, 2002). Zur Erläuterung enthält [Tabelle 1](#) beispielhaft für 15 Befragte die Befragungsmerkmale Alter sowie erreichte Punktzahl bei einem Test.

**Tabelle 1**  
Beispielhafte Erhebung der Merkmale Alter und erreichte Punktzahl für 15 Befragte

| Befragte Person | Alter in Jahren | Punktzahl |      |     |      |
|-----------------|-----------------|-----------|------|-----|------|
|                 |                 | insgesamt | MCAR | MAR | MNAR |
| 1               | 18              | 13        | –    | 13  | 13   |
| 2               | 21              | 6         | 6    | –   | –    |
| 3               | 23              | 5         | 5    | –   | –    |
| 4               | 25              | 19        | 19   | –   | 19   |
| 5               | 32              | 8         | –    | –   | –    |
| 6               | 38              | 14        | 14   | 14  | 14   |
| 7               | 40              | 11        | 11   | 11  | 11   |
| 8               | 43              | 15        | 15   | 15  | 15   |
| 9               | 48              | 8         | –    | 8   | –    |
| 10              | 51              | 12        | 12   | 12  | 12   |
| 11              | 54              | 14        | 14   | 14  | 14   |
| 12              | 58              | 18        | –    | 18  | 18   |
| 13              | 66              | 9         | 9    | 9   | –    |
| 14              | 72              | 18        | 18   | 18  | 18   |
| 15              | 85              | 15        | –    | 15  | 15   |

MCAR: missing completely at random; MAR: missing at random; MNAR: missing not at random.

4 Zur Kompensation dieses Ausfalls werden häufig Gewichtungsverfahren genutzt (Kalton/Kasprzk, 1986; Brick/Kalton, 1996; Brick, 2013; Gabler/Ganninger, 2010). Jedoch finden sich auch Ansätze zum Einsatz von Imputationsverfahren im Rahmen von Unit Nonresponse (Messingschlager, 2012).

Wichtig ist nun, worin der Datenausfall begründet liegt. Im einfachsten Fall ist die Wahrscheinlichkeit eines Datenausfalls unabhängig von den beobachteten und unbeobachteten Daten – und damit rein zufällig. In solch einem Fall spricht man von einem Missing-Completely-At-Random-(MCAR-)Ausfall. Übertragen auf das Beispiel in Tabelle 1 (Spalte „MCAR“) fallen die Werte für die Punktzahl rein zufällig aus.<sup>15</sup>

Komplizierter wird es, wenn ein systematischer Ausfall vorliegt. Kann der Ausfall dabei vollständig auf die beobachteten Daten bedingt werden, liegt ein Missing-At-Random-(MAR-)Ausfall vor. In der Spalte „MAR“ fallen die Werte für die Punktzahl insbesondere für diejenigen Befragten mit einem geringen Alter aus.<sup>16</sup> Somit kann die Wahrscheinlichkeit für einen Ausfall der Punktzahl mithilfe der beobachteten Werte des Alters erklärt werden. Innerhalb der jeweiligen Altersgruppen kann nun erneut ein rein zufälliger (MCAR-)Ausfall für das Merkmal Punktzahl angenommen werden.

Dies ist nicht mehr gegeben, wenn der Ausfall eines Merkmals durch sich selbst bedingt ist. In einem solchen Fall spricht man von einem Missing-Not-At-Random-(MNAR-)Ausfall. In der Spalte „MNAR“ fallen lediglich die niedrigen Punktzahlen aus. Die Ausprägung des Merkmals selbst verursacht hier also dessen Ausfall. Somit ergibt sich ein Problem, da die benötigte Information darüber, weshalb ein Wert fehlt, nicht in den beobachteten Daten zu finden ist.<sup>17</sup> Hier ist es notwendig, auf externe Informationsquellen zuzugreifen beziehungsweise spezifische Annahmen für den Ausfall zu treffen (Raghunathan, 2015, hier: Kapitel 7; Hammon/Zinn, 2020).<sup>18,19</sup>

- 5 Dies würde es erlauben, eine durchschnittliche Punktzahl auf Grundlage der beobachteten Werte unverzerrt zu schätzen.
- 6 Im obigen Beispiel liegt sowohl eine positive Korrelation zwischen dem Alter und der Ausfallwahrscheinlichkeit der Punktzahl als auch zwischen dem Alter und der Punktzahl selbst vor. Ein Durchschnittswert, der sich lediglich auf den beobachteten Daten bemisst, überschätzt die Punktzahl hier entsprechend tendenziell.
- 7 Welcher Ausfallmechanismus im konkreten Fall vorliegt, kann oft nicht allein aus den Daten heraus abgeleitet werden. Vielmehr ist hierbei eine Einschätzung durch Fachstatistikerinnen und Fachstatistiker entscheidend, gegebenenfalls unterstützt durch Heranziehung externer Informationen.
- 8 Das Alter kann auch hier – zumindest erwartungsgemäß – einen Teil des Ausfalls erklären (siehe Fußnote 6 zum Zusammenhang zwischen dem Alter und der [Ausfallwahrscheinlichkeit für die] Punktzahl), jedoch bleibt hier ein unerklärter „MNAR-Rest“.
- 9 Die Behandlung eines MNAR-Ausfalls ist im weiteren Verlauf nicht Gegenstand dieses Artikels.

Unter der Annahme, dass am Ende der späteren Analyse lediglich Aussagen über Populationsparameter (zum Beispiel eine Summe oder ein arithmetisches Mittel) getroffen werden sollen (was ja auch der Normalfall ist), so gilt: Ist der Ausfall MCAR oder MAR, vereinfacht dies die Spezifikation des Behandlungsmodells, anhand dessen letztlich valide Parameterschätzungen beziehungsweise -inferenzen erzielt werden sollen, erheblich.

Nachdem dargestellt wurde, in welchen Fällen fehlende Werte ein Problem für die weitere Analyse darstellen, stellt sich nun die Frage, welche Ansätze die Literatur – auch neben der Imputation – bietet, um die fehlenden Werte zu behandeln.

## 3

### Alternativen im Umgang mit fehlenden Werten

Im Fall von maximum-likelihood-basierten Schätzern von Populationskenngrößen kann gegebenenfalls direkt mit dem unvollständigen Datenmaterial gearbeitet werden, indem die eigentliche Likelihoodfunktion über alle möglichen Werte der fehlenden Daten integriert wird (Little/Rubin 2002, Kapitel 8). Andernfalls lautet die sicherlich einfachste Antwort, Beobachtungen, die fehlende Werte in ihren Merkmalen aufweisen, aus der weiteren Analyse auszuschließen (complete case analysis, listwise deletion). Ein solches Vorgehen jedoch verschwendet erhoebene Informationen, indem sie nicht genutzt werden. Zudem besteht je nach Ausfallmechanismus die Möglichkeit, dass die aus dem so reduzierten Datenmaterial geschätzten Größen eine unerwünschte Verzerrung aufweisen (siehe Kapitel 2). Ähnliches gilt, wenn jeweils nur diejenigen Beobachtungen herangezogen werden, die für die Berechnung einer spezifischen Kennzahl relevant und verfügbar sind (pairwise deletion).<sup>10</sup> Der einfachste Weg ist also nicht unbedingt der beste.

Alternativ können fehlende Werte durch plausible Werte ersetzt werden (Imputation). Der Datensatz wird auf diese Weise also vervollständigt.

<sup>10</sup> Zum Beispiel würden für das Merkmal Punktzahl lediglich die beobachteten Daten dieses Merkmals, für eine weitere Variable hingegen nur die beobachteten Daten dieser weiteren Variable verwendet. Die Fallzahlen unterscheiden sich demnach kennzahl-spezifisch.

## 4

### Methoden zur Imputation

Im Folgenden werden verschiedene Ansätze zur Imputation vorgestellt. Dabei liegt das Augenmerk auf solchen Ansätzen, die die Imputation ausschließlich anhand des vorhandenen Datenmaterials vornehmen. Das Hinzuziehen externer Quellen (unter anderem direktes Recherchieren fehlender Werte) als weitere Möglichkeit wird hier nicht weiter thematisiert.

Für eine Imputation stehen mehrere Verfahren zur Verfügung. Die Wahl des Verfahrens ist abhängig von der Datensituation, vom Ausfall (bezüglich dessen Ausmaßes und Mechanismus) und vom konkreten Imputationsziel zu treffen. Insbesondere gibt es nicht eine beste Methode, die bei allen Anwendungsfällen immer die besten Ergebnisse erzielt. Neben der deduktiven Einsetzung, die – sofern möglich – vor der eigentlichen Imputation stattfindet, sind bei der Imputation zwei grundlegende Herangehensweisen zu unterscheiden: Bei der Single-Imputation wird für einen fehlenden oder unplausiblen Wert genau ein Wert imputiert. Dieses Vorgehen birgt jedoch die Gefahr, sich in falscher Sicherheit zu wiegen: So werden, falls keine Varianz korrigierenden Verfahren zum Einsatz kommen (Bruch, 2016), imputierte Werte während der späteren Datenanalyse als solche behandelt, die von Anfang an beobachtet waren. Dieser Unsicherheit, insbesondere bei der Berechnung von relativen Standardfehlern – und damit bei der Inferenz von der Stichprobe auf die interessierende Grundgesamtheit –, ist angemessene Rechnung zu tragen. Bei der zweiten Herangehensweise, der multiplen Imputation, wird ein fehlender Wert hingegen mehrfach ersetzt.

#### 4.1 Deduktive Imputation

Zu Beginn der Konzeption eines Imputationsvorgehens ist zu prüfen, ob es logische Zusammenhänge zwischen den Merkmalen gibt, die nur eine deduktive Einsetzung zulassen. Zum Beispiel lässt sich der Familienstand einer dreijährigen Person nur durch die Einsetzung „ledig“ sinnvoll befüllen.<sup>11</sup>

<sup>11</sup> Hinter deduktiven Einsetzungen liegt die Annahme, dass die beobachteten Werte (also im genannten Beispiel das Alter) korrekt sind.

## 4.2 Single-Imputation

Es gibt zahlreiche Verfahren, die zu einer Single-Imputation führen. Diese lassen sich grob nach univariaten und multivariaten Verfahren unterscheiden.

### Univariate Imputationsverfahren

Univariate Methoden nutzen lediglich Informationen, die das jeweilige Merkmal selbst bietet. Zusammenhänge zu weiteren Merkmalen, wie im Falle eines MAR-Ausfallmechanismus (siehe Kapitel 2) eigentlich gegeben, werden dagegen nicht berücksichtigt. Ein einfaches, univariates Imputationsverfahren ist die Imputation basierend auf Lagemaßen. Dabei erfolgt die Einsetzung zum Beispiel durch das arithmetische Mittel, sodass alle fehlenden Werte eines Merkmals durch den gleichen Wert imputiert werden.<sup>12</sup> Ein weiteres, einfaches Verfahren ist die Einsetzung durch einen zufälligen Wert desselben Merkmals einer anderen Beobachtungseinheit, die Random-Hot-Deck-Imputation. Liegen zum Beispiel bei Zeitreihen Vorerhebungswerte vor, werden häufig Methoden wie „last observation carried forward“ und „baseline observation carried forward“ (van Buuren, 2018) angewendet. Dabei wird der letzte beobachtete Wert übernommen<sup>13</sup>, oder ein geeignetes Mittel vergangener Werte derselben Beobachtungseinheit genutzt<sup>14</sup>.

### Multivariate Imputationsverfahren

#### *Regressionsbasierte Imputationsverfahren*

Um, in Anlehnung an die Ausfallmechanismen aus Kapitel 2, Zusammenhänge zwischen Merkmalen bei der Imputation zu nutzen, werden zur Schätzung der fehlenden Werte multivariate Verfahren verwendet. Häufig handelt es sich hierbei um Regressionsansätze (einschließlich solcher zur Schätzung von kategorialen Merkmalen). Bei diesen Verfahren wird anhand der Einheiten mit vollständig beobachteten (oder iterativ vervollständigten) Werten ein Modell geschätzt, welches anschließend

für die Schätzung der unbeobachteten Werte verwendet wird. Die gewählten unabhängigen Variablen des Modells sollten sowohl das zu imputierende Merkmal als auch den Ausfall erklären können (siehe Beispiel zum MAR-Ausfallmechanismus aus Kapitel 2). Zudem sollte das Modell die Merkmale umfassen, die zusammen mit dem zu imputierenden Merkmal für die spätere Ergebniserstellung ausgewertet werden. Dabei ist es zunächst unerheblich, welche konkrete Methode zur Regression eingesetzt wird. Infrage kommen neben der klassischen linearen Regression mit ihren Erweiterungen auch Splines, additive Modelle, baumbasierte Verfahren, Support Vector Machines und viele andere mehr.<sup>15</sup>

Schätzungen auf Basis solcher Methoden können dazu führen, dass die Zusammenhangsmaße (zum Beispiel die Korrelation) in den imputierten Daten und somit auch im gesamten Datenbestand überschätzt werden. Ursache dafür ist, dass alle Schätzungen auf der erstellten Vorhersagefunktion (Hyperebene im linearen Fall, etwas Gestuftes bei baumbasierten Verfahren, etwas Geschwungenes bei Support Vector Machines und so weiter) liegen. Um diesem Problem entgegenzuwirken, kann den imputierten Werten eine zusätzliche Unsicherheit in Form einer stochastischen Komponente, zum Beispiel ein zufällig gezogenes Residuum aus der Verteilung der beobachteten Residuen, hinzugefügt werden (van Buuren, 2018).

#### *Spenderbasierte Imputationsverfahren*

Eine weitere Gruppe von Imputationsverfahren sind spenderbasierte Verfahren, bei denen für einen fehlenden Wert ein beobachteter Wert einer anderen Beobachtungseinheit eingesetzt wird. Die Regressionsimputation kann zu einem spenderbasierten Verfahren modifiziert werden, indem als Imputation dann nicht die durch das Modell ermittelte Vorhersage verwendet wird. Stattdessen wird diese Vorhersage mit den Vorhersagen der vollständig beobachteten Einheiten verglichen. Eine der  $k$  ( $k \geq 1$ ) Beobachtungen, deren Vorhersagen ähnlich sind, kann dann als Spender verwendet werden. Diese Methode wird als Predictive Mean Matching bezeichnet (zur Anwendung siehe Lange/Pöttsch, 2019). Auch für Entscheidungsbäume existiert ein Ansatz, der den spenderbasierten Verfahren zugeordnet wird (Pech, 2019).

12 Zwar wird dann der Mittelwert der Stichprobe bei einem MCAR-Ausfall trotz der Imputation erwartungstreu geschätzt (und zwar nur für diesen Ausfallmechanismus); die Varianz wird jedoch unterschätzt, da alle fehlenden Zellen durch den gleichen Wert, und damit „zu sicher“, imputiert werden.

13 Hinter diesem Vorgehen steckt die Annahme, dass sich der Wert in der Zwischenzeit nicht wesentlich verändert hat.

14 Dieses Vorgehen zählt zu den sogenannten Cold-Deck-Verfahren, bei denen externe Quellen für die Imputation hinzugezogen werden.

15 Ein Anwendungsbeispiel bietet Schreiner/Schmidt (2011), eine grundlegende Einführung in verschiedene Regressionsmethoden James und andere (2013).

Schließlich gehören auch die klassischen Nearest-Neighbour-Verfahren (Nächste-Nachbar-Verfahren) zu den spenderbasierten Methoden (Baumgärtner und andere, 2018). Die Idee ist hier, einen fehlenden Wert durch einen beobachteten Wert einer anderen, möglichst ähnlichen Spendereinheit zu ersetzen. Im Gegensatz zum Predictive Mean Matching, bei dem die Ähnlichkeit hinsichtlich der Vorhersagen des zu imputierenden Merkmals betrachtet wird, erfolgt die Ähnlichkeitssuche bei einem Nearest-Neighbour-Verfahren über Ähnlichkeits- oder Distanzmaße bezüglich der erklärenden Merkmale.

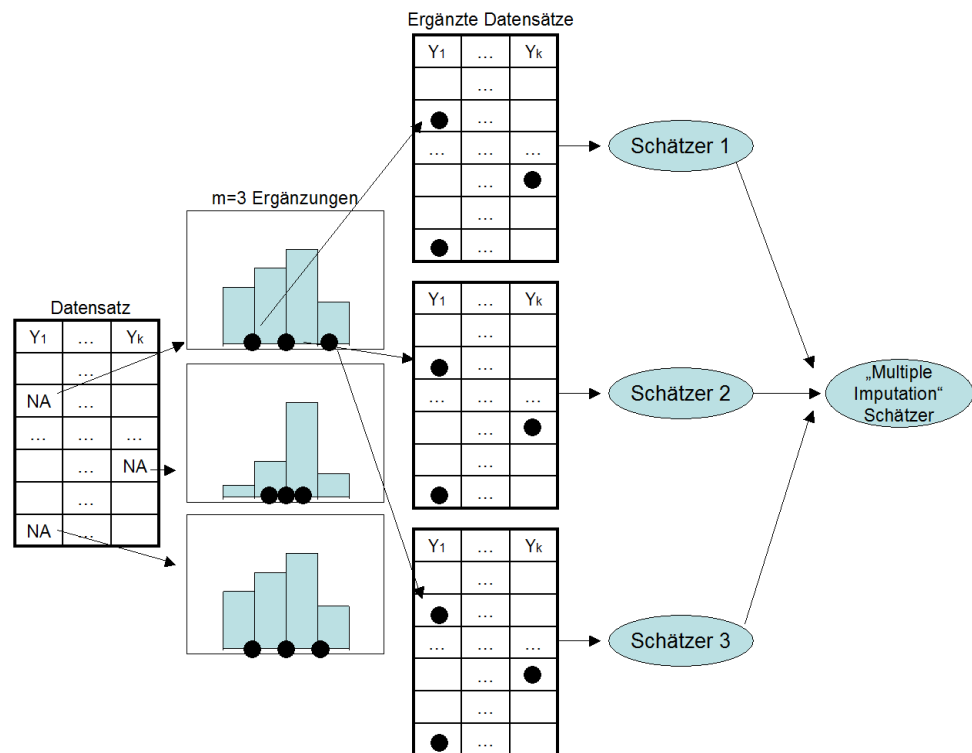
## 4.3 Multiple Imputation

Informationen, die in einem Datenbestand nicht vorliegen, können auch mit dem besten Verfahren keine sicheren Größen werden. Imputationen sind daher stets unsicher. Single-Imputation unterschlägt diese Tatsache, sofern keine Varianz korrigierenden Verfahren zum Einsatz kommen. Durch die Gleichstellung der zuvor

unbeobachteten Werte mit den beobachteten innerhalb der Analyse wird eine zu große Sicherheit der imputierten Werte suggeriert. Da anhand von auf so vervollständigten Daten basierenden Stichprobenschätzern Rückschlüsse auf die interessierende Grundgesamtheit gezogen werden sollen, stellt sich die Frage, wie diese Unsicherheit berücksichtigt werden kann. Ein Ansatz dazu ist die multiple Imputation.

Der multiplen Imputation (Rubin, 1987; Rubin, 1996; Murray, 2018) zugrunde liegt die Idee, fehlende Daten nicht durch jeweils einen imputierten Wert zu ersetzen, sondern durch mehrere. Dies führt zur Frage, woher mehrere (verschiedene) imputierte Werte kommen können. Üblicherweise (wenngleich nicht notwendigerweise) wird die bayesianische Interpretation von Wahrscheinlichkeit genutzt. Sie besagt, dass unbekannte Größen – zum Beispiel die fehlenden Werte selbst oder auch die Koeffizienten bei einem linearen Regressionsansatz – einer Wahrscheinlichkeitsverteilung folgen. Diese kann auf Basis der im Datensatz vorhandenen Informationen als A-posteriori-Verteilung geschätzt wer-

**Grafik 1**  
Das Prinzip der multiplen Imputation



Quelle: Rässler/Meinfelder (2011)

2021 - 0361

den. Bei mehrfachen zufälligen Ziehungen aus dieser erhält man mehrere Imputationen und damit mehrere imputierte Datensätze, auf deren Basis die interessierenden Populationsparameter geschätzt werden können. Dieses Vorgehen führt zu mehreren Schätzwerten je Populationsparameter, worin sich die durch die Imputation eingebrachte Unsicherheit widerspiegelt.

#### ↳ Grafik 1

Als finale Schätzung für einen Populationsparameter auf Basis mehrerer imputierter Datensätze wird klassisch gemäß der combining rules von Rubin (1987) schließlich das arithmetische Mittel derjenigen Schätzungen herangezogen, die sich bei Verwendung von jeweils nur einem imputierten Datensatz ergeben. Eine etwas aufwändigere Formel ergibt sich für die Schätzung der Varianz der so durchgeführten Schätzung des Populationsparameters.<sup>16</sup> Hinreichend große Fallzahlen erlauben darüber hinaus auch Aussagen über die approximative Verteilung der Schätzer und darauf aufbauend die Durchführung von Tests. Inwiefern gültige Inferenz auf Basis multipel imputierter Datensätze möglich ist, ist im jeweiligen Einzelfall genau zu prüfen.

Die multiple Imputation bietet für die statistischen Ämter bei der Bereitstellung von Daten Vorteile aus wissenschaftlicher Sicht: Ein statistisches Amt kann – neben der Durchführung eigener Schätzungen – mit einem kleinen Satz von vollständigen Datensätzen approximativ gültige Inferenzen für eine Vielzahl von potenziellen Analysen auch außerhalb des Hauses ermöglichen. Alle Analysen, die später auf den dann publizierten Daten beruhen, können auf demselben Satz von Imputationen basieren. Dieses Vorgehen stellt sicher, dass Unterschiede in den Ergebnissen nicht auf den Umgang mit fehlenden Daten zurückzuführen sind. Dieses Vorgehen ist aus wissenschaftlicher Sicht einem Vorgehen vorzuziehen, bei dem lediglich angegeben wird, welche Werte imputiert wurden.

---

16 Zu Rubins Ansatz existieren Alternativen (Aßmann/Preising, 2020).

## 5

---

### Besondere Herausforderungen in der Praxis

---

Die in Kapitel 4 vorgestellten Methoden liefern sozusagen das Handwerkszeug für die Imputation fehlender Werte. In der Praxis können allerdings abhängig von der Datensituation Anpassungen notwendig sein. Hierfür einige Beispiele: Der Zusammenhang zwischen zu imputierendem Merkmal und Prädiktorvariablen kann durchaus komplex sein. Hier kann eine Transformation (zum Beispiel Logarithmieren) Abhilfe schaffen (van Buuren, 2018). Merkmale mit einer bimodalen Verteilung können mit einem zweistufigen Verfahren imputiert werden. Bei Merkmalen wie Anteilen oder Summen, die sich aus anderen ableiten lassen, ist es sinnvoll, nur einen Teil zu imputieren und die anderen aus den dann zur Verfügung stehenden Werten abzuleiten.

Merkmale nacheinander in einer bestimmten Reihenfolge zu imputieren kann bei der Einhaltung allgemeiner Plausibilitätsregeln (zum Beispiel Minima oder Maxima für einzelne Merkmale, Anforderungen an Kombinationen von auftretenden Merkmalsausprägungen) nützlich sein. Insbesondere in der amtlichen Statistik ist dies häufig von Bedeutung. So ist es möglich, Zusammenhänge zwischen den Merkmalen durch multivariate Verfahren auf Basis einer Regression zu berücksichtigen. Alternativ kann die Einhaltung solcher Regeln teilweise mit der Verwendung eines Nearest-Neighbour-Verfahrens sichergestellt werden.<sup>17</sup> Außerdem kann auch die Verwendung von Imputationsklassen<sup>18</sup> dazu beitragen, plausible Merkmalskombinationen zu generieren. Sofern die Schätzungen der Populationsparameter beziehungsweise die Momente oder die Verteilung der imputierten Daten hinreichende Qualität aufweisen, können Verletzungen der Plausibilität auf Einzeldatenebene je nach Verwendungszweck auch toleriert werden (Kleinke und andere, 2020).

---

17 Das Werkzeug CANCEIS verfolgt diesen Ansatz.

18 Imputationsklassen unterteilen die Daten in homogene Untergruppen. Die Zugehörigkeit zu einer Imputationsklasse definiert sich über ein oder mehrere Merkmale, deren Ausprägungen bei allen Datensätzen dieser Imputationsklasse übereinstimmen oder hinreichend ähnlich sein müssen.



Bei großen Datenmengen und für den Einsatz komplexer Methoden und Algorithmen ist es aus praxisorientierter Sicht zudem notwendig, eine geeignete IT-Infrastruktur bereitzustellen. Abhängig von der Größe der Datenbestände müssen ausreichende Speicherkapazitäten und Rechenleistungen bereitstehen, um die Verfahren automatisiert durchführen zu können. Andernfalls ist mit langen Laufzeiten zu rechnen.

Eine weitere Herausforderung in der Praxis besteht darin, dass die Qualität der Schätzungen immer auch von der Qualität der beobachteten Daten abhängt, da diese in der Regel als Prädiktoren in die Modellbildung oder als Spender in die Imputation einfließen. Aus diesem Grund ist sicherzustellen, dass diese Hilfsinformationen weitgehend fehlerfrei sind.

Bei der Imputation sind die wahren Werte gerade unbekannt. Daher sind geeignete Testdaten und/oder Annahmen nötig, um das beste Imputationsverfahren auswählen zu können. Wie eine Evaluation des Imputationsverfahrens gelingen kann, auch ohne die unbekanntenen Werte zu kennen, erläutert das folgende Kapitel.

## 6

### Verfahrensauswahl und -evaluation

Bei der Auswahl von maschinellen Imputationsverfahren, beim Produktivbetrieb und bei der Bewertung der Imputationsergebnisse für jeweils konkrete Anwendungsfälle sind vielfältige Aspekte zu berücksichtigen. Erforderlich ist deswegen eine enge Zusammenarbeit von Fachstatistik, Methodik und IT-Bereich.

#### 6.1 Auswahl, Spezifikation und Test des Imputationsverfahrens

Muss in einem konkreten Anwendungsfall untersucht werden, welches Imputationsverfahren (mit welchen Spezifikationen) zum Einsatz kommen soll, sind unter anderem Testläufe unerlässlich, um die Imputationsgüten der infrage kommenden Verfahren einschätzen zu können. Dabei sind auch Annahmen über den Ausfallmechanismus und das Ausmaß des Ausfalls zu treffen. Das übliche Vorgehen ist wie folgt: Ausgehend von einem vollständigen und plausiblen, das heißt fehler-

freien Datensatz werden innerhalb einer Simulationsstudie fehlende Werte gemäß der unterschiedlichen, für die jeweilige Fragestellung relevanten Ausfallmechanismen künstlich erzeugt. Anschließend werden die zu testenden Imputationsverfahren auf den künstlich unvollständigen Datensatz angewandt. Zu diesem Zeitpunkt liegt also sowohl der wahre, vollständige Datensatz vor als auch je Imputationsverfahren ein oder mehrere vervollständigte Datensätze. Deren Abweichungen voneinander können je nach Imputationsziel in geeigneter Weise berechnet werden. Kleine Abweichungen sprechen dabei für eine hohe Imputationsgüte, große Abweichungen für eine geringe Imputationsgüte.

Die Ziele von Imputation – und damit die Metriken zur Bestimmung der Abweichungen – können dabei zwar divergieren<sup>19</sup>; grundsätzlich ist das Ziel von Imputation jedoch eine erwartungstreue und bezüglich der Inferenz valide<sup>20</sup> Schätzung von Populationsparametern (van Buuren, 2018, hier: Kapitel 2.5.2). Zu den dazugehörigen Metriken zählen in der Regel die Verzerrung sowie Varianz der Stichprobenschätzer und gegebenenfalls auf Konfidenzbändern basierende Überdeckungsrate(n) (sogenannte Coverages). Sind Subgruppen oder Schichten als besonders relevant identifiziert, sollten – neben den globalen Parametern – auch die subgruppen- beziehungsweise schichtspezifischen Parameterschätzer explizit untersucht werden.<sup>21</sup> Diese Metriken können numerisch-statistisch bestimmt und ausgewertet werden. Darüber hinaus sind aber auch grafische Analysen nützlich.

19 Geht es darum, für einzelne statistische Einheiten möglichst exakte Angaben zu schätzen – Ziel: predictive accuracy; zum Beispiel zur Erstellung vollständiger, korrekter Register (Di Zio und andere, 2020) –, so spricht die Literatur zum Teil ebenfalls von Imputation (Chambers, 2006).

20 Das heißt eine Inferenz, die sowohl die Stichprobenunsicherheit als auch die Unsicherheit aufgrund der fehlenden Werte berücksichtigt.

21 Eventuell ist zum Zeitpunkt der Imputation allerdings (noch) nicht bekannt, welche Parameterschätzer anhand des vervollständigten Datensatzes im Zentrum der künftigen Analyse stehen. Auch erscheinen gegebenenfalls zu viele Parameterschätzer als gleichermaßen hoch relevant. In solchen Fällen sollte das Ziel darin bestehen, dass sich zumindest die niederen (gegebenenfalls auch gemischten) Momente einer univariaten (gegebenenfalls auch multivariaten) Merkmalsverteilung nicht signifikant von denjenigen vor der Löschung unterscheiden (estimation accuracy) oder, weitergedacht, die Verteilungen selbst keine signifikanten Unterschiede aufweisen (distributional accuracy). Die Untersuchung hierfür sinnvoll anwendbarer Metriken sind Teil aktueller Arbeiten (Thurrow und andere, 2021).

Weitere wichtige Aspekte sind unter anderem die benötigte Laufzeit und der benötigte Arbeitsspeicher eines Verfahrens für die Imputation, die grundsätzliche Eignung von Verfahren für die vorliegenden Merkmalskassen sowie das Vorliegen verlässlicher Implementierungen.

## 6.2 Produktivbetrieb und Ergebnisbewertung

---

Das Vorgehen, bei dem die interessierenden Parameterschätzer vor der eigenhändigen Löschung bekannt sind und als Vergleichsgröße herangezogen werden können, ist im Produktivbetrieb, also innerhalb einer laufenden Statistikproduktion, offensichtlich nicht anwendbar: Schließlich liegen die hierfür theoretisch benötigten vollständig plausiblen Daten eben nicht vor, sondern müssen zunächst unter Zuhilfenahme des Imputationsverfahrens geschätzt werden. Es ist somit nicht möglich, mit quantifizierbarer Sicherheit zu entscheiden, wie erfolgreich die Imputation erfolgt ist. Falls verfügbar, können externe Daten beziehungsweise Veröffentlichungen zur Validierung der imputierten Daten herangezogen werden. In jedem Fall bedarf es Analysen mithilfe von Grafiken, Kennzahlen oder statistischen Tests sowohl durch Methodikerinnen und Methodiker als auch durch die Fachstatistik (Nguyen und andere, 2017; Bondarenko/Raghunathan, 2016).

Grundsätzlich gilt hierfür: Imputierte Daten sollten so aussehen, als wären sie beobachtet worden (van Buuren, 2018; de Waal und andere, 2011). Das heißt nicht, dass die imputierten Daten dieselben Verteilungsmomente besitzen müssen wie die beobachteten: Fehlen für ein Merkmal insbesondere kleine Werte (siehe Tabelle 1, in der bei „MAR“ und „MNAR“ tendenziell die geringeren Punktzahlen fehlen), so sollte zum Beispiel das arithmetische Mittel der imputierten Daten auch kleiner sein als das der beobachteten. Somit sollte ein Vergleich zwischen den imputierten und beobachteten Werten immer vor dem Hintergrund erfolgen, dass ein wohlspezifiziertes Imputationsmodell gegebenenfalls gerade dafür sorgt, dass sich beide Verteilungen voneinander unterscheiden. Gleichzeitig gilt es, Zusammenhänge der im Datensatz befindlichen Merkmale zu erhalten: Korrelieren zwei Merkmale in den gemeinsam beobachteten Daten, so sollte dies auch im imputierten Teil der Fall sein. Zusammengefasst sollten die imputier-

ten Werte sowohl univariate als auch multivariate Spezifika der beobachteten Daten abdecken beziehungsweise erhalten.<sup>22</sup>

## 7

---

### Dokumentation

---

Im Anschluss an die Imputation ist eine sorgfältige Dokumentation unerlässlich. Neben den imputierten Daten sollten auch die Rohdaten und entsprechende Qualitätskennzeichen (Luzi und andere, 2008) gespeichert werden. Diese geben mindestens an, ob und in welchem Schritt ein Wert imputiert wurde, und sind im Aufbereitungsprogramm mit zu implementieren. Nur mit vorhandenen Rohdaten sind methodische Untersuchungen zu Auswirkungen der Imputation auf die statistischen Ergebnisse möglich. Auch erlauben sie es, den Einfluss des Imputationsprozesses auf unplausible Werte und Merkmalskombinationen, die vorab im Rohmaterial enthalten waren und plausibilisiert wurden, zu testen.

Zur Dokumentation gehört auch eine Beschreibung des Imputationsvorgehens und dessen Auswirkungen. Van Buuren (2018) liefert Empfehlungen für zu dokumentierende Inhalte bei der multiplen Imputation. Diese können größtenteils auch bei Single Imputation verwendet werden. Beispiele, wie sich solche Angaben in einem Text zusammensetzen lassen, der auch an Leserinnen und Leser adressiert ist, die sich mit Imputation nicht auskennen, finden sich in Enders (2010).

Ziel ist es, den Imputationsprozess nachvollziehbar zu machen. Dadurch ist auch im Nachhinein eine Einschätzung der Verwendungsbreite der imputierten Daten möglich. Für spezifische Auswertungen (zum Beispiel durch externe Wissenschaftlerinnen und Wissenschaftler), bei denen andere Annahmen, Ziele oder zusätzliche Hilfsmerkmale zugrunde liegen, können so alternative Imputationen durchgeführt werden. Dies bietet auch die Möglichkeit, eine nachgelagerte multiple Imputation

---

22 Weitere Ausführungen zur Identifikation und Beschreibung von fachstatistischen und methodischen Qualitätsindikatoren, die zur Evaluation eines Plausibilisierungs- beziehungsweise Imputationsverfahrens im Zuge eines laufenden Statistikproduktionsprozesses im Echtbetrieb herangezogen werden können, enthält ein internes Dokument (auf Anfrage bei der Autorenschaft erhältlich).

durchzuführen, auf deren Basis dann auch die Standardfehler, die die Unsicherheit der Imputation mit einbeziehen, unverzerrt geschätzt werden können.

### 8

---

## Fazit und Empfehlungen


---

Plausibilisierung und Imputation spielen eine wichtige Rolle im Produktionsprozess der amtlichen Statistik. Aktuelle Beispiele sind Untersuchungen zur Anwendung eines Nearest-Neighbour-Verfahrens zur Behandlung fehlender und fehlerhafter Werte im Zensus 2022 und bei der monatlichen Verdiensterhebung. Ein weiteres Beispiel ist die Imputation nicht erhobener Merkmale bei kleinen Unternehmen in der Strukturerhebung im Dienstleistungsbereich mit einem zweistufigen regressionsbasierten Ansatz.

Insgesamt führen

- › die Spezifika und die Vor- und Nachteile der einzelnen in diesem Aufsatz skizzierten Klassen von Imputationsverfahren,
- › die Bedeutung des Ausfallmechanismus,
- › der Einfluss des Ziels der Imputation sowie
- › die Notwendigkeit, Imputationsergebnisse zu bewerten,

zu folgenden Empfehlungen:

Bei allen Fragen der Imputation, insbesondere bei der Neukonzeption von Imputationsschritten in der Datenaufbereitung einer Statistik sind eine enge Zusammenarbeit der Bereiche Fachstatistik und Methodik sowie die in Abschnitt 6.2 beschriebene fortwährende Evaluation dringend angeraten. So kann der Imputationsprozess zu einer hohen Qualität der Ergebnisse amtlicher Statistik beitragen. 

## LITERATURVERZEICHNIS

---

- Aßmann, Christian/Preising, Marcel. *Bayesian estimation and model comparison for linear dynamic panel models with missing values*. In: Australian & New Zealand Journal of Statistics. Jahrgang 62. Ausgabe 4/2020, Seite 536 ff.  
DOI: [onlinelibrary.wiley.com](https://onlinelibrary.wiley.com)
- Baumgärtner, Luisa/Gräb, Christopher/Leppert, Philipp/Söllner, René/Spies, Lydia/Veith, Stefan/Vorgrimler, Daniel. *Imputation und Konsolidierung: Neue Aufgaben für die Unternehmensstatistik*. In: WISTA Wirtschaft und Statistik. Ausgabe 6/2018, Seite 33 ff.
- Bondarenko, Irina/Raghunathan, Trivellore. *Graphical and numerical diagnostic tools to assess suitability of multiple imputations and imputation models*. In: Statistics in Medicine. Jahrgang 35. Ausgabe 17/2016, Seite 3007 ff. DOI: [10.1002/sim.6926](https://doi.org/10.1002/sim.6926)
- Brick, J. Michael. *Unit Nonresponse and Weighting Adjustments: A Critical Review*. In: Journal of Official Statistics. Jahrgang 29. Ausgabe 3/2013, Seite 329 ff.  
DOI: [10.2478/jos-2013-0026](https://doi.org/10.2478/jos-2013-0026)
- Brick, J. Michael/Kalton, Graham. *Handling missing data in survey research*. In: Statistical Methods in Medical Research. Jahrgang 5. Ausgabe 3/1996, Seite 215 ff.  
DOI: [10.1177/096228029600500302](https://doi.org/10.1177/096228029600500302)
- Bruch, Christian. *Varianzschätzung unter Imputation und bei komplexen Stichproben-  
designs*. Trier 2016.
- Carpenter, James R./Kenward, Michael G. *Multiple Imputation and its Application*. Chichester 2013.
- Chambers, Ray. *Evaluation criteria for editing and imputation in Euredit*. In: United Nations Statistical Commission and United Nations Economic Commission for Europe (Herausgeber). Statistical Data Editing. Jahrgang 3. New York, Genf 2006. Seite 17 ff.
- de Waal, Ton/Pannekoek, Jeroen/Scholtus, Sander. *Handbook of Statistical Data Editing and Imputation*. New York 2011.
- Di Zio, Marco/Filippini, Romina/Rocchetti, Gaia. *An imputation procedure for the Italian attained level of education in the register of individuals based on administrative and survey data*. 2020. [Zugriff am 23. August 2021]. Verfügbar unter: [www.unecce.org](http://www.unecce.org)
- Enders, Craig K. *Applied missing data analysis*. New York 2010.
- Esser, Hartmut/Grohmann, Heinz/Müller, Walter/Schäffer, Karl-August. *Mikrozensus im Wandel: Untersuchungen und Empfehlungen zur inhaltlichen und methodischen Gestaltung*. In: Statistisches Bundesamt (Herausgeber). Band 11 der Schriftenreihe Forum der Bundesstatistik. Stuttgart 1989.
- Gabler, Siegfried/Ganninger, Matthias. *Gewichtung*. In: Wolf, Christof/Best, Henning (Herausgeber). Handbuch der Sozialwissenschaftlichen Datenanalyse. Wiesbaden 2010. Seite 143 ff.

## LITERATURVERZEICHNIS

---

Gehle, Christian/Lüüs, Hans-Peter. *Prozessmanagement im Statistischen Bundesamt*. In: WISTA Wirtschaft und Statistik. Ausgabe 5/2017, Seite 46 ff.

Grunwald, Sven/Krause, Anja. *Umgang mit fehlenden Angaben in der Gebäude- und Wohnungszählung 2011*. In: Wirtschaft und Statistik. Ausgabe 8/2014, Seite 437 ff.

Hammon, Angelina/Zinn, Sabine. *Multiple imputation of binary multilevel missing not at random data*. In: Journal of the Royal Statistical Society: Series C (Applied Statistics). Jahrgang 69. Ausgabe 3/2020, Seite 547 ff. DOI: <https://doi.org>

James, Gareth/Witten, Daniela/Hastie, Trevor/Tibshirani, Robert. *An Introduction to Statistical Learning*. New York 2013.

Kalton, Graham/Kasprzok, Daniel. *The Treatment of Missing Survey Data*. In: Survey Methodology. Jahrgang 12. Heft 1/1986, Seite 1 ff.

Kleinke, Kristian/Reinecke, Jost/Salfrán, Daniel/Spiess, Martin. *Applied Multiple Imputation*. Cham 2020.

Lange, Kerstin/Pöttsch, Olga. *Neues Imputationsverfahren bei Antwortausfällen zu geborenen Kindern im Mikrozensus*. In: WISTA Wirtschaft und Statistik. Ausgabe 5/2019, Seite 9 ff.

Little, Roderick J. A./Rubin, Donald B. *Statistical Analysis with Missing Data*. 2. Auflage. Hoboken, New Jersey 2002.

Luzi, Orietta/Di Zio, Marco/Guarnera, Ugo/Manzari, Antonia/de Waal, Ton/Pannekoek, Jeroen/Hoogland, Jeffrey/Tempelman, Caren/Hulliger, Beat/Kilchmann, Daniel. *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*. 2008. [Zugriff am 23. August 2021]. Verfügbar unter: [ec.europa.eu](http://ec.europa.eu)

Messingschlager, Martin. *Fehlende Werte in den Sozialwissenschaften – Analyse und Korrektur mit Beispielen aus dem ALLBUS*. Bamberg 2012.

Murray, Jared S. *Multiple Imputation: A Review of Practical and Theoretical Findings*. In: Statistical Science. Jahrgang 33. Ausgabe 2/2018, Seite 142 ff.

Nguyen, Cattram D./Carlin, John B./Lee, Katherine J. *Model checking in multiple imputation: an overview and case study*. In: Emerging Themes in Epidemiology. Jahrgang 14. Ausgabe 8/2017, Seite 1 ff.

Pech, Birgit. *Maschinelles Lernen: Classification and Regression Trees (CART) für die Imputation nutzbar machen*. In: Zeitschrift für amtliche Statistik Berlin Brandenburg. Ausgabe 3/2019, Seite 12 ff.

Raghuathan, Trivellore. *Missing Data Analysis in Practice*. New York 2015.

Rässler, Susanne/Meinfelder, Florian. *Behandlung fehlender Werte in Erhebungen*. Unveröffentlichtes Skript. Lehrstuhl für Statistik und Ökonometrie. Universität Bamberg. 2011.

## LITERATURVERZEICHNIS

---

Rubin, Donald B. *Multiple imputation after 18+ years*. In: Journal of the American Statistical Association. Jahrgang 91. Ausgabe 434/1996, Seite 473 ff.

Rubin, Donald B. *Multiple Imputation for Nonresponse in Surveys*. New York 1987.

Rubin, Donald B. *Multiple Imputations in Sample Surveys – A Phenomenological Bayesian Approach to Nonresponse*. In: Proceedings of the Survey Research Method Section of the American Statistical Association. 1978, Seite 20 ff.

Schreiner, Carsten/Schmidt, Katrin. [Nacherhebung Bewässerung zur Landwirtschaftszählung 2010](#). In: Wirtschaft und Statistik. Ausgabe 12/2011, Seite 1202 ff.

Statistische Ämter des Bundes und der Länder. *Qualitätshandbuch der Statistischen Ämter des Bundes und der Länder*. Wiesbaden 2021. [Zugriff am 23. August 2021]. Verfügbar unter: [www.destatis.de](http://www.destatis.de)

Thurow, Maria/Dumpert, Florian/Ramosaj, Burim/Pauly, Markus. *Goodness (of fit) of Imputation Accuracy: The GoodImpact Analysis*. 2021. [Zugriff am 23. August 2021]. Verfügbar unter: <https://arxiv.org>

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

**Herausgeber**  
Statistisches Bundesamt (Destatis), Wiesbaden

---

**Schriftleitung**  
Dr. Daniel Vorgrimler  
Redaktion: Ellen Römer

---

**Ihr Kontakt zu uns**  
[www.destatis.de/kontakt](http://www.destatis.de/kontakt)

---

**Erscheinungsfolge**  
zweimonatlich, erschienen im Oktober 2021  
Ältere Ausgaben finden Sie unter [www.destatis.de](http://www.destatis.de) sowie in der [Statistischen Bibliothek](#).

---

Artikelnummer: 1010200-21005-4, ISSN 1619-2907

---

© Statistisches Bundesamt (Destatis), 2021  
Vervielfältigung und Verbreitung, auch auszugsweise, mit Quellenangabe gestattet.