Scharfe, Simone; Racky, Matthias; Lange, Kerstin

Article

Fehlende Datensätze in der Meldung: Imputation versus Neuanforderung. Möglichkeiten und Grenzen der Imputation bei der Erhebung zu Tarifinformationen

WISTA – Wirtschaft und Statistik

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

Suggested Citation: Scharfe, Simone; Racky, Matthias; Lange, Kerstin (2024) : Fehlende Datensätze in der Meldung: Imputation versus Neuanforderung. Möglichkeiten und Grenzen der Imputation bei der Erhebung zu Tarifinformationen, WISTA – Wirtschaft und Statistik, ISSN 1619-2907, Statistisches Bundesamt (Destatis), Wiesbaden, Vol. 76, Iss. 1, pp. 39-53

This Version is available at:
https://hdl.handle.net/10419/284653

FEHLENDE DATENSÄTZE IN DER MELDUNG: IMPUTATION VERSUS NEUANFORDERUNG

Möglichkeiten und Grenzen der Imputation bei der Erhebung zu Tarifinformationen

Simone Scharfe, Matthias Racky, Kerstin Lange

↘ **Schlüsselwörter:** Imputation – Simulation – Verdienststatistikgesetz – Nearest-Neighbor-Prinzip – Belastungsreduzierung

ZUSAMMENFASSUNG

Die monatliche Verdiensterhebung hat die vierjährlich durchgeführte Verdienststrukturerhebung und die Vierteljährliche Verdiensterhebung abgelöst. Ergänzt wird sie durch eine fünfjährliche Erhebung zum angewandten Tarifvertrag. Es kommt vor, dass Betriebe nicht zu allen Arbeitnehmersätzen der monatlichen Verdiensterhebung Angaben liefern, die Meldung also unvollständig ist. Die hier vorgestellte Studie untersucht, ob es möglich ist, die Angaben zur tarifvertraglichen Bindung durch ein Nearest-Neighbor-Imputationsverfahren mit einer hinreichenden Zuverlässigkeit für das Gesamtergebnis zu schätzen.

↘ ***Keywords:*** *imputation – simulation – Earnings Statistics Act – nearest-neighbor principle – burden reduction*

ABSTRACT

The monthly earnings survey has replaced the four-yearly structure of earnings survey and the quarterly earnings survey. It is supplemented by a five-yearly survey of the collective agreements applied. Sometimes the reporting companies do not provide information on the relevant collective agreement for all employee records in the monthly earnings survey, and so the report is incomplete. The study presented here examines whether a nearest-neighbor imputation can be used to estimate the information on collective bargaining coverage while ensuring adequate reliability of the overall results.

Simone Scharfe

ist Diplom-Kauffrau und Diplom-Handelslehrerin und leitet das Referat „Arbeitskostenerhebung, Tarifstatistiken“ des Statistischen Bundesamtes. Ihre Tätigkeitsschwerpunkte liegen in der konzeptionellen und methodischen Weiterentwicklung sowie Datenanalyse der Tarif- und Arbeitskostenerhebungen.

Matthias Racky

ist im Referat „Arbeitskostenerhebung, Tarifstatistiken“ des Statistischen Bundesamtes für die Vorbereitung und Durchführung der Arbeitskostenerhebung, die Aktualisierung und Gewährleistung der Ergebnisqualität des Arbeitskostenindex sowie differenzierte Auswertungen zur Tarifbindung zuständig.

Kerstin Lange

hat Statistik studiert und ist Referentin im Referat „Künstliche Intelligenz, Big Data“ des Statistischen Bundesamtes. Sie entwickelt Verfahren zur Imputation und automatisierten Plausibilisierung, unter anderem aus dem Bereich des maschinellen Lernens. Zudem berät sie bei deren Implementierung in die Statistikproduktion.

1

Einleitung

Mit der letzten Änderung des Verdienststatistikgesetzes im Jahr 2020 wurden die vierjährlich bei Betrieben durchgeführte Verdienststrukturerhebung und die Vierteljährliche Verdiensterhebung ab dem Jahr 2022 durch die monatliche Verdiensterhebung abgelöst (Finke und andere, 2023). Alle fünf Jahre ergänzt eine Erhebung zum angewandten Tarifvertrag beim jeweiligen Arbeitnehmer beziehungsweise der jeweiligen Arbeitnehmerin die neue Verdiensterhebung.

Die fünfjährliche Erhebung zum Tarifvertrag erfolgt aus erhebungstechnischen Gründen im Nachgang der Erhebung zu den monatlichen Verdiensten. Die Verknüpfung der Datensätze der beiden Erhebungen geschieht über das Hilfsmerkmal der Personalnummer. Gleichwohl kann es sein, dass im Zuge der Tariferhebung von Betrieben nicht zu allen Arbeitnehmersätzen der monatlichen Verdiensterhebung Angaben zum Tarifvertrag geliefert werden, die Meldung also unvollständig ist. Nach der klassischen Methode müsste in diesem Fall der meldende Betrieb zu einer nochmaligen (vollständigen) Meldung aufgefordert werden. In der im Folgenden vorgestellten Simulationsstudie wird untersucht, ob (und wenn ja, bis zu welchem Anteil von fehlenden Arbeitnehmersätzen im Betrieb) die Angaben zur tarifvertraglichen Bindung durch ein Nearest-Neighbor-Imputationsverfahren mit einer hinreichenden Zuverlässigkeit für das Gesamtergebnis geschätzt werden können. Die Belastung der Meldenden und der Statistischen Ämter der Länder soll reduziert werden, gleichzeitig ist die Genauigkeit der Ergebnisse sicherzustellen.

Die Umstellung der Verdiensterhebungen infolge der Anpassung des Verdienststatistikgesetzes sowie das Ziel der Simulationsstudie sind Thema in Kapitel 2. In Kapitel 3 wird das Untersuchungsdesign ausführlich dargestellt, die Ergebnisse der Imputations-Simulationsstudie enthält Kapitel 4. Ein Ausblick beschließt den Artikel.

2

Hintergründe und Ziel der Studie

2.1 Umstellung der Verdiensterhebungen

Die letzte Änderung des Verdienststatistikgesetzes im August 2020 führte zu einem grundlegenden Wandel der Datenerhebung in den Verdienststatistiken. Die vierteljährliche Verdienststatistik bei rund 40 500 Betrieben und die vierjährliche Verdienststrukturstatistik bei rund 60 000 Betrieben wurden in einer monatlichen Erhebung zusammengeführt. Sie ermittelt Angaben zu Verdiensten und Arbeitszeiten sowie zu betrieblichen, arbeitsplatzbezogenen und persönlichen Charakteristika zu jedem Arbeitnehmer und jeder Arbeitnehmerin im Betrieb. Die meisten dieser Angaben sind in den betrieblichen Lohnabrechnungssystemen vorhanden und können im Idealfall direkt aus den Lohnabrechnungssystemen digital an die Statistischen Ämter der Länder übermittelt werden.

Dieses neue Erhebungsdesign gewährleistet zum einen die Analysepotenziale der Vergangenheit. Zum anderen liegen Daten nun zu politisch und gesellschaftlich hochinteressanten Fragestellungen, beispielsweise zum bereinigten Gender Pay Gap, zu Effekten der Mindestlohnerhöhung oder auch zum Niedriglohn, in einer jährlichen statt vierjährlichen Periodizität vor (Finke und andere, 2023). ↘ Übersicht 1

Von der monatlichen Lieferung ausgenommen ist die Information, ob und nach welchem Tarifvertrag der einzelne Arbeitnehmer beziehungsweise die einzelne Arbeitnehmerin entlohnt wird. Zwei Überlegungen führten im Rahmen des Gesetzgebungsverfahrens dazu, hierfür eine separate Erhebung vorzusehen:

1. Die Ergebnisse dieser Erhebung werden zentral für das Wägungsschema des Tarifindex[1] benötigt. Dieses ist in einem fünfjährlichen Rhythmus anzupassen und erfordert daher keine jährliche oder gar monatliche Aktualisierung.

1 Der Tarifindex des Statistischen Bundesamtes ist nicht nur ein wichtiger Indikator zur Beobachtung der konjunkturellen Entwicklung, sondern spielt auch eine zentrale Rolle bei der Entscheidung über eine Anpassung des gesetzlichen Mindestlohnes (siehe § 9 Mindestlohngesetz).

Übersicht 1

Gegenüberstellung des alten und des neuen Erhebungskonzeptes in den Verdienststatistiken

Bisheriges Konzept	Neues Konzept nach der Änderung des Verdienststatistikgesetzes 2020
Verdienststrukturerhebung › vierjährliche Stichprobenerhebung bei rund 60 000 Betrieben › Angaben zu Verdiensten, Arbeitszeiten › Angaben zur Entlohnung nach (welchem) Tarifvertrag › auf Arbeitnehmerebene › zusätzliche Strukturmerkmale › mit Unterstichprobe › zuletzt 2018	Verdiensterhebung › monatliche Stichprobenerhebung bei rund 58 000 Betrieben › Angaben zu Verdiensten, Arbeitszeiten › zusätzliche Strukturmerkmale › für jeden Arbeitnehmer/jede Arbeitnehmerin mit Angabe der Personalnummer › erstmals: Januar 2022
	↓ Schlüsselkriterien: Personalnummer und BerichtseinheitsID
Vierteljährliche Verdiensterhebung › vierteljährliche Stichprobenerhebung bei rund 40 500 Betrieben › zu durchschnittlichen Verdienst-/Arbeitsangaben für Beschäftigtengruppen › zuletzt durchgeführt für das 4. Quartal 2021	Erhebung zu Tarifinformationen › fünfjährliche Erhebung als Unterstichprobe aus Verdiensterhebung von maximal 20 000 Betrieben › Merkmal: angewandter Tarifvertrag › als Eingliederungsnummer › für jeden Arbeitnehmer/jede Arbeitnehmerin mit Angabe der Personalnummer › erstmals für September 2025
Analysepotenzial	
› unterjährige Verdienstentwicklung › alle 4 Jahre Aussagen zu: › Verdienstverteilung › Verdiensten nach Berufen › bereinigtem Gender Pay Gap, Niedriglohn-, Mindestlohnanalyse › Tarifbindung › Wägungsschema für Tarifindex	› unterjährige Verdienstentwicklung › jährliche Aussagen zu: › Verdienstverteilung › Verdiensten nach Berufen › bereinigtem Gender Pay Gap, Niedriglohn-, Mindestlohnanalyse › alle 5 Jahre: › Wägungsschema für Tarifindex › Paneldaten

2. Die Angaben zum jeweils angewandten Tarifvertrag liegen im erforderlichen Lieferformat in der Regel nicht in der Lohnsoftware der Meldebetriebe vor. Eine Erhebung ist damit in einer monatlichen Taktung mit sehr kurzen Lieferfristen nicht realisierbar.

Für eine einheitlich verwertbare Meldung, die an die Informationen der Tarifdatenbank des Statistischen Bundesamtes gekoppelt werden kann, meldet der Betrieb den angewandten Tarifvertrag je Arbeitnehmer/-in mit der sogenannten Eingliederungsnummer. Diese recherchiert der Meldebetrieb wiederum in der Tarifdatenbank oder er wendet sich an das zuständige Statistische Landesamt, wenn er sie nicht findet. Insbesondere bei Firmentarifverträgen ist es durchaus möglich, dass diese noch nicht in der Tarifdatenbank eingegliedert sind.

Aus den vorgenannten Gründen werden die (für den Tarifindex notwendigen) Tarifinformationen separat erhoben. Um eine Doppelerhebung von Angaben zu vermeiden, ist diese separate Erhebung zu Tarifinformationen (ETI) in zwei Punkten an die monatliche Verdiensterhebung (VE) gekoppelt:

1. Die Meldebetriebe für die ETI sind eine Unterstichprobe der Betriebe aus der VE, die im Berichtsmonat der ETI gemeldet haben.

2. Angaben zu Verdiensten, Arbeitszeiten und weiteren Strukturmerkmalen werden in der ETI nicht nochmals erhoben, sondern über die Hilfsmerkmale Personalnummer[2] und BerichtseinheitsID[3] des Meldebetriebs aus der Meldung zur VE an die Ergebnisse der ETI angespielt.

Dieses Vorgehen hilft, Redundanzen in Erhebungen zu den Verdiensten zu vermeiden.

2 Die Personalnummer stellt hierbei ein Hilfsmerkmal der statistischen Erhebung dar und wird nach Erstellung der Ergebnisdatei gelöscht.

3 BerichtseinheitsID = Kennnummer des Meldebetriebes.

↘ Ergebnisse der Erhebung zu Tarifinformationen und das Wägungsschema des Tarifindex[4]

Mit den Ergebnissen der ETI liegen künftig für jede Branche detaillierte Informationen zur Anwendung von Tarifverträgen vor. Die Besonderheit und somit der große Vorteil für die Berechnung der Tarifindizes liegt darin, dass die ETI Informationen auf Ebene der einzelnen Arbeitnehmerinnen und Arbeitnehmer zur Verfügung stellen kann. Nur die Angaben zum Tarifverdienst der tatsächlich nach Tarifvertrag entlohnten Arbeitnehmerinnen und Arbeitnehmer fließen in das Wägungsschema der Tarifindizes ein. Für jede Abteilung (Zweisteller) der Klassifikation der Wirtschaftszweige (Statistisches Bundesamt, 2008) werden für Deutschland insgesamt, für das frühere Bundesgebiet sowie für die neuen Länder in der Regel jeweils so viele Tarifverträge in das Wägungsschema der Tarifindizes aufgenommen, bis mindestens 75 % aller Beschäftigten, die nach Tarifverträgen bezahlt werden, abgedeckt sind. Um dabei die Anzahl an Tarifverträgen in Grenzen zu halten, werden vornehmlich die Tarifverträge mit der höchsten Anzahl an Tarifbeschäftigten ausgewählt. Dies können sowohl Branchen- als auch Firmentarifverträge sein. Die oben genannte 75-%-Schranke stellt eine Minimumgrenze dar, die je nach Wirtschaftszweig variiert. Im Bereich „Öffentliche Verwaltung, (...)“ ist beispielsweise mit einem Abdeckungsgrad von 100 % eine Totalerfassung garantiert. Damit solche Bereiche wegen ihres hohen Abdeckungsgrads nicht überproportional in die gesamtwirtschaftlichen Indizes einfließen, werden die ausgewählten Tarifverträge auf Ebene der übrigen Branchen so hochgerechnet, dass sie zahlenmäßig ebenfalls alle nach Tarifvertrag bezahlten Arbeitnehmerinnen und Arbeitnehmer der jeweiligen Branche repräsentieren.

2.2 Ziel der Simulationsstudie

Wie in Abschnitt 2.1 geschildert, sollen die für den Berichtsmonat September 2025 in der ETI gewonnenen Informationen zum angewandten Tarif (als Eingliederungsnummer) mit den für jede Beschäftigte und jeden Beschäftigten erhobenen Daten aus der VE des gleichen Berichtsmonats über die Personalnummer des Meldebetriebes verknüpft werden. Damit kommt der Personalnummer als Schlüsselvariable besondere Bedeutung zu.

Der Meldebetrieb wird aufgefordert, zur ETI für den identischen Personenkreis mit der identischen Personalnummer zu melden wie für die VE im September 2025. Gleichwohl lässt die Praxis vermuten, dass es hier zu Differenzen durch die Verwendung veränderter Personalnummernsystematiken, aber auch eines anderen gemeldeten Personenkreises kommen kann. Szenarien der gemeldeten Personalnummern in den beiden Erhebungen aus einem Betrieb sind in ↘ Übersicht 2 schematisch dargestellt.

Während in Szenario I die gemeldete Personalnummernstruktur beider Meldungen übereinstimmt und die Daten beider Erhebungen problemlos zusammengeführt werden können, sind in den weiteren Szenarien jeweils Unterschiede zu entdecken.

In Szenario II fehlen Angaben zur Tarifinformation für die Arbeitnehmer M und N. Diese sollten jedoch noch durch die ETI-Erhebung vervollständigt werden. Klassisch wäre hier, dass der Betrieb noch einmal angeschrieben wird. Im Zuge der in diesem Artikel beschriebenen Simulationsstudie soll geprüft werden, ob die Informationen für die Arbeitnehmer M und N alternativ über eine Imputation anhand der gemeldeten Tarifinformationen für die anderen Arbeitnehmer dieses Betriebes geschätzt werden können. Im zweiten Schritt soll geprüft werden, welche Ausfallquote je Betrieb für die Erhebung vertretbar ist.

Aus Sicht der ETI stellt sich Szenario III relativ unproblematisch dar: Die zusätzlich in der ETI gemeldeten Datensätze können gelöscht, die restlichen problemlos mit der VE zusammengeführt werden. Bei Szenario IV ist schnell klar, dass mit dieser Meldung nicht weitergearbeitet werden kann, sondern der Meldebetrieb kontaktiert und um eine neue Meldung für die ETI gebeten werden muss.

Szenario V ist letztlich eine Sonderform von Szenario II beziehungsweise kann ausschließlich über eine Imputation oder die Hochrechnung gelöst werden. Hintergrund dieses Falles könnte beispielsweise sein, dass der Meldebetrieb zwei Teillieferungen für zwei Betriebsteile in der VE liefert, beide aber jeweils mit der gleichen Nummernsystematik unterlegt. Damit kommt es zu Dubletten der Personalnummer innerhalb eines Betriebes. Während sich die VE mit dem Anspielen eines differenzierenden Merkmals (zum Beispiel dem Geburtsjahr) behelfen kann, besteht in der ETI diese Möglichkeit nicht, da außer der Eingliederungsnummer kein weiteres Merkmal in der Meldung zur Verfügung steht. Hier bleibt letztlich nur die Möglichkeit, die doppelt geliefer-

4 Weitere Informationen enthält der Methodenbericht zum Index der Tarifverdienste (Statistisches Bundesamt, 2021).

Übersicht 2
Szenarien der Meldedatensätze Verdiensterhebung (VE) und Erhebung zu Tarifinformationen (ETI) aus einem Meldebetrieb

Szenario I		Szenario II		Szenario III		Szenario IV		Szenario V	
identische Meldestruktur		fehlende Angaben in ETI		zusätzliche Personen in ETI		andere Personalnummern-systematik		Dubletten in Meldung	
VE	ETI	VE	ETI	VE	ETI	VE	ETI	VE	ETI
PersNr.	PersNr.	PersNr.	PersNr.	PersNr.	PersNr.	PersNr.	PersNr.	PersNr.	PersNr.
A	A	A	A	A	A	C1	A	A	A
B	B	B	B		B	C2	B	B	B
C	C	C	C	C	C	C3	C	C	C
D	D	D	D	D	D	C4	D	D	D
E	E	E	E	E	E	C5	E	E	E
F	F	F	F	F	F	C6	F	F	F
G	G	G	G	G	G	C7	G	G	G
H	H	H	H	H	H	C8	H	H	H
I	I	I	I	I	I	C9	I	I	I
J	J	J	J		J	D1	J	J	J
K	K	K	K	K	K	D2	K	K	K
L	L	L	L	L	L	D3	L	A	A
M	M	M		M	M	D4	M	B	B
N	N	N		N	N	D5	N	C	C

ten Eingliederungsnummern zu löschen und anschließend durch das Imputationsergebnis zu ersetzen.

Ziel der nachfolgend dargestellten Analyse ist zu prüfen,

- ob und in welcher Spezifikation eine Imputation hinreichend treffsichere Ergebnisse erzielt, damit sich der Aufwand einer Neuanforderung der Meldung beim Betrieb erübrigt, sowie
- welcher Anteil an fehlenden Informationen zum angewandten Tarif des Betriebes für die ETI tolerierbar ist.

Die Bewertung der Ergebnisse der Imputationen erfolgt in diesem Beitrag anhand von Imputationsfehlerquoten. Weiterführende Analysen mit Blick auf die (zentralen) statistischen Ergebnisse sind Teil nachfolgender Untersuchungen (Kapitel 5).

3 Untersuchungsdesign

3.1 Der Testdatensatz

Die Verdienststrukturerhebung 2018 bildet die Datenbasis der Untersuchung. Hier liegen vollständige Informationen zum angewandten Tarifvertrag auf Arbeitnehmerebene vor. Die rund 60 000 auf Basis einer geschichteten Stichprobe ausgewählten Betriebe hatten verpflichtend zu melden, nach welchem Tarifvertrag der jeweilige Arbeitnehmer beziehungsweise die jeweilige Arbeitnehmerin entlohnt wurde. Lag dem Arbeitsvertrag kein Tarifvertrag zugrunde, war auch dies zu melden.

Für die im Folgenden beschriebene Imputationssimulationsstudie wurde der Datensatz der Verdienststrukturerhebung 2018 wie folgt eingegrenzt:

- Es werden ausschließlich von Betrieben gemeldete Datensätze verwendet, also keine aus sekundären Daten abgeleitete Datensätze.[5]
- Es werden ausschließlich Betriebe mit mindestens einem Arbeitnehmerdatensatz mit Tarifinformation[6] ausgewählt.
- Es werden keine Arbeitnehmersätze ausgeschlossen, das heißt es sind auch Datensätze zu Auszubildenden, Praktikanten und Beschäftigten in Altersteilzeit enthalten.
- Es werden ausschließlich Betriebe mit mehr als einem Arbeitnehmerdatensatz berücksichtigt.

Insgesamt enthielt der Testdatensatz damit 12 956 Betriebe und 292 093 Arbeitnehmerdatensätze mit 1 633 unterschiedlichen angewandten Tarifverträgen (Eingliederungsnummern).

↘ Tabelle 1 verdeutlicht, dass der überwiegende Teil der Betriebe einen beziehungsweise zwei angewandte Tarifverträge gemeldet hat. Rund 7 % der Betriebe meldeten fünf (oder mehr) unterschiedliche Tarifverträge.

Tabelle 1
Strukturmerkmale hinsichtlich Tarifinformationen

	Betriebe	
	Anzahl	%
1 angewandter Tarifvertrag im Betrieb	5 417	41,8
2 angewandte Tarifverträge im Betrieb	4 246	32,8
3 angewandte Tarifverträge im Betrieb	1 700	13,1
4 angewandte Tarifverträge im Betrieb	717	5,5
5 (und gegebenenfalls mehr) angewandte Tarifverträge im Betrieb	876	6,8

An dieser Stelle sei darauf hingewiesen, dass die Meldung des Betriebes auch Arbeitnehmerinnen und Arbeitnehmer enthalten kann, die nicht nach Tarifvertrag entlohnt werden.

5 Nähere Erläuterungen enthält der Qualitätsbericht zur Verdienststrukturerhebung (Statistisches Bundesamt, 2018).

6 Diese Einschränkung ergibt sich daraus, dass nur für diese Meldebetriebe künftig eine potenzielle Imputation erforderlich sein wird.

3.2 Der Simulationsansatz

Der Simulationsansatz der Studie leitet sich aus dem Grundgedanken ab, bei einer Imputation fehlender Angaben zum angewandten Tarifvertrag Datensätze mit gelieferten Angaben aus dem gleichen Betrieb zu verwenden. Die Angaben zum angewandten Tarifvertrag werden bei einigen Datensätzen mittels Zufallsverfahren gelöscht, um dann durch Imputation neu generiert zu werden. Anschließend erfolgt eine Gegenüberstellung der imputierten mit den realen Angaben bei verschiedenen Ausfallquoten.

Als Startsimulation wird in jedem der 12 956 Betriebe entsprechend des Simulationslaufes für

10, 15, 20, 25, 30 beziehungsweise 50 %

der Arbeitnehmerdatensätze die Angabe zum angewandten Tarifvertrag zufällig gelöscht. Da die zufällige Löschung auf Betriebsebene erfolgt, kann der Anteil der fehlenden Angaben im Gesamtdatensatz von der vorgegebenen Quote abweichen. Insgesamt wurden für jedes Szenario jeweils 15 Simulationsläufe erstellt. Über die Standardabweichung der Ergebnisse ist es möglich, im Anschluss die Stabilität der Ergebnisse zu bewerten (siehe Kapitel 4).

3.3 Der Imputationsansatz

Nearest-Neighbor-Imputationsverfahren

In der Vergangenheit konnten in der amtlichen Statistik unter anderem sowohl in den Verdienststatistiken (Frentzen/Günther, 2017; Schymura, 2020) als auch in den Unternehmensstatistiken (Preising und andere, 2021) durch den Einsatz von Imputationsverfahren fehlende Angaben in Meldungen kompensiert werden. Wie in diesen Ansätzen soll auch hier das Nearest-Neighbor-Imputationsverfahren auf Anwendbarkeit getestet werden.

Die Nearest-Neighbor-Methode zählt zu den Hot-Deck-Techniken (Little/Rubin, 2002, hier: Seite 69). Grundsätzlich wird hierbei für fehlende Werte eines Empfängerdatensatzes aus demselben Basisdatenmaterial nach Spendern gesucht. Für den hier vorliegenden Ansatz wird dies noch weiter eingeschränkt: Es soll ausschließlich nach Spendern im gleichen Betrieb gesucht wer-

den. Damit finden der Aspekt der betriebsspezifischen Anwendung von Firmentarifverträgen und Betriebsvereinbarungen, aber auch der ausgewählte Einsatz (entsprechend Branchen/Region) von Branchentarifverträgen besondere Berücksichtigung in der Modellierung.

Die Spender-Arbeitnehmerdatensätze sollten eine möglichst ähnliche Charakteristik wie die jeweiligen Empfänger-Arbeitnehmerdatensätze aufweisen. Es wird der Spender gesucht und ausgewählt, der eine möglichst geringe Distanz zum Empfänger aufweist. Die Distanz wird dabei numerisch durch eine sogenannte Distanzfunktion berechnet.

Imputation über CANCEIS

Diese Studie verwendete für die Imputationen die Software CANCEIS (Canadian Census Edit and Imputation System). Sie wurde 1992 vom kanadischen Statistikamt Statistics Canada/Statistique Canada für die Aufbereitung und Imputation von Zensusdaten entwickelt und in den letzten Jahren so weit angepasst, dass sie heute auch für andere Erhebungen international genutzt wird (CANCEIS Development Team, 2015). Bei dem in CANCEIS implementierten Algorithmus handelt es sich um einen Nearest-Neighbor-Ansatz.

Für die Berechnung der Distanz D_{fp} zwischen dem ursprünglichen Empfänger und dem Spender werden zunächst Distanzen D_i für jede Variable i berechnet. Diese ergeben sich aus der Anwendung der in Abhängigkeit des Skalenniveaus gewählten Distanzfunktion auf die Daten und sind auf den Bereich zwischen 0 (Merkmalswert des Spenders ist identisch zum Empfängerwert) und 1 (bei kategorialen Merkmalen: Merkmalswert des Spenders unterscheidet sich vom Empfängerwert; bei numerischen Merkmalen: Merkmalswert des Spenders unterscheidet sich stark vom Empfängerwert) normiert. Um die Gesamtdistanz daraus zu berechnen, werden diese Einzeldistanzen mit dem jeweiligen Gewicht des Merkmals w_i multipliziert und über alle Merkmale aufsummiert, also $D_{fp} = \sum_i w_i D_i$. Diese Distanz bestimmt dann die nächsten Nachbarn, von denen einer als Spender fungiert.

Spezifikation

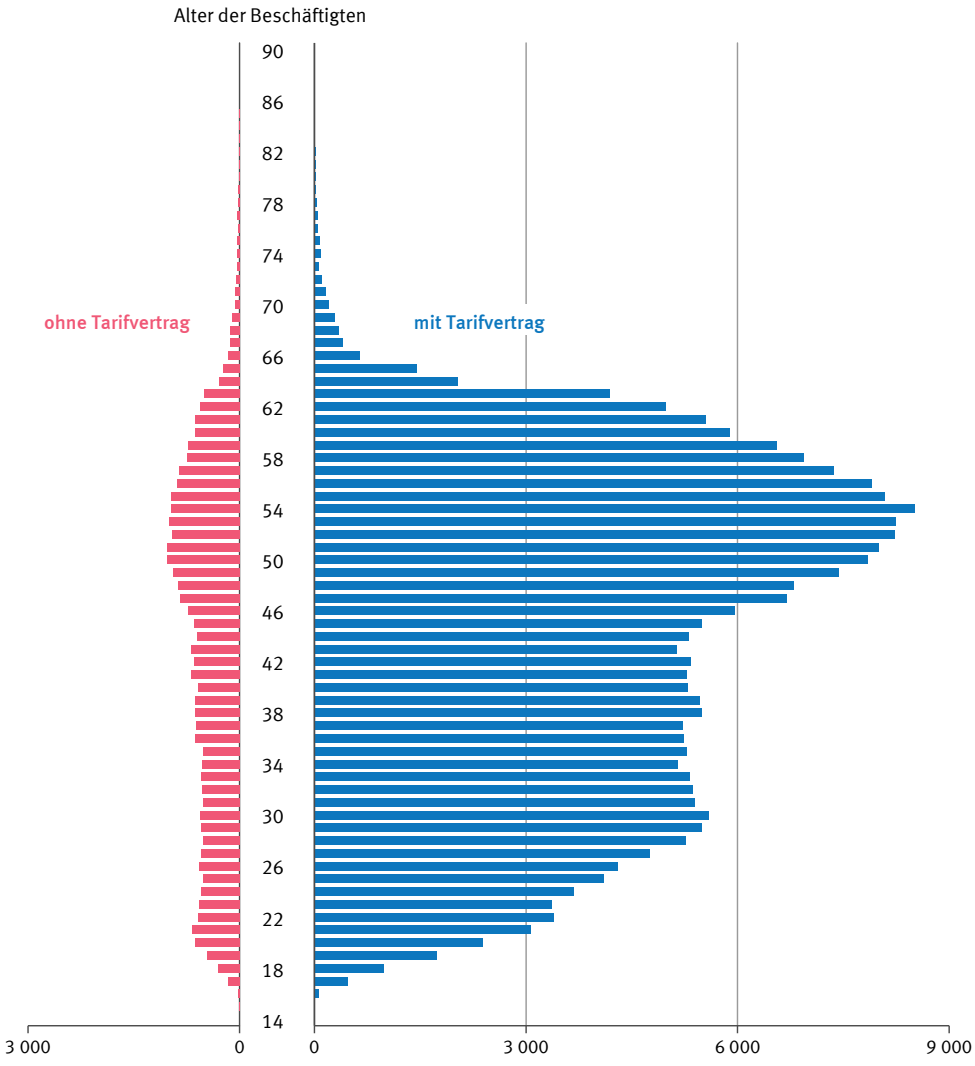
Mit Blick auf den künftigen Einsatz für die ETI können für die Imputation nur die Variablen der Verdienststrukturerhebung 2018 einbezogen werden, die auch im Datensatz der VE vorhanden und künftig mit der ETI verknüpfbar sein werden. Eine erste deskriptive Analyse für Arbeitnehmerinnen und Arbeitnehmer mit und ohne Tarifvertrag in Betrieben, in denen mindestens ein Arbeitnehmer beziehungsweise eine Arbeitnehmerin anhand eines Tarifvertrages entlohnt wurde, wies darauf hin, dass

- Beschäftigte mit einem Tarifvertrag durchschnittlich einen höheren Bruttostundenverdienst (über alle Größenklassen) realisieren konnten und
- es marginal unterschiedliche Personengruppen-Strukturen in der Gruppe der Beschäftigten mit und ohne Tarifvertrag gibt. ↘ Grafik 1 auf Seite 46

Diese deskriptive Analyse stellt jedoch nur die Gruppen mit und ohne Tarifvertrag einander gegenüber. In der Imputation geht es hingegen insbesondere auch darum, die richtige Eingliederungsnummer beziehungsweise zumindest eine Eingliederungsnummer, die es in dem Betrieb gibt, zuzuordnen.

Für eine möglichst zielgenaue Imputation der Eingliederungsnummer soll aus dem eigenen Betrieb imputiert werden. Daher wird dieser Variable ein deutlich höheres Gewicht (=100) als den anderen Erklärvariablen zugewiesen (=1). Dies stellt sicher, dass immer ein Spender aus demselben Betrieb verwendet wird, solange es dort mindestens einen potenziellen Spender gibt. Basis der Spezifikation der Distanzfunktionen bilden insbesondere Erfahrungen aus früheren Anwendungen von CANCEIS in den Verdienststatistiken, insbesondere bei den Verdiensterhebungen 2015 bis 2019 sowie der Verdienststrukturerhebung.

Grafik 1
Beschäftigte mit und ohne Tarifvertrag in Betrieben, in denen mindestens ein Tarifvertrag angewendet wird
Deskriptive Analyse auf Basis der Verdienststrukturerhebung 2018



4

Ergebnisse

4.1 Einblick in einzelne CANCEIS-Reports

CANCEIS gibt neben dem Einzeldatenmaterial, in dem die fehlenden Eingliederungsnummern mit der Imputation ergänzt sind, auch die in ↘ Tabelle 2 zusammengefassten Übersichten mit Metainformationen zum Erfolg des CANCEIS-Laufs aus.

Das Beispiel stellt den ersten CANCEIS-Lauf aus dem Szenario eines (simulierten) 20-prozentigen Ausfalls beim Merkmal Eingliederungsnummer dar.

Von den 292 093 Datensätzen (Beschäftigten) im Testdatensatz sind 63 569 (21,76 %) durch eine fehlende Information zur Eingliederungsnummer gekennzeichnet. Für alle diese Datensätze konnte ein (hinreichend) passender Spender gefunden und die Imputation erfolgreich durchgeführt werden. Wie gut diese Imputation den „wahren Wert“ abbildet, erläutert Abschnitt 4.2. Insgesamt 54 622 Spenderdatensätze wurden herangezogen. Einige Spender kamen somit mehrfach zur Anwen-

Tabelle 2
CANCEIS-Imputationsergebnisübersicht

Classification of Edited Units

	Datensätze	
	Anzahl	%
Total Units Edited	292 093	100
Passed Units	228 524	78,24
Failed Units	63 569	21,76
Total Units Processed	292 093	100
Passed Units	228 524	78,24
Failed Units	63 569	21,76
Containing Only Invalid Values	63 569	21,76
Containing Only Inconsistent Values	0	0,00
Containing Both Invalid and Inconsistent Values	0	0,00
Successfully Imputed	63 569	21,76
Donor Units Chosen During Imputation	54 622	18,70
Passing Inconsistency Edits	54 622	18,70
Failed Inconsistency Edits	0	0,00
Passing Outlier Edits	54 622	18,70
Failed Outlier Edits	0	0,00
Failed Imputation	0	0,00
No NMCIAs found	0	0,00
Other reasons	0	0,00
Units Not Processed	0	0,00
Passed Units	0	0,00
Failed Units	0	0,00

Title	Statistics on Imputed Data
Module	ETI
Stratum	1
Generated On	2023-08-16 9:33
CANCEIS Version	5.2.1423.1420

Distance Distributions of Imputed Units

Percentile	D_{fp}
0	0
1	0
5	0,02
10	0,04
25	0,14
50	0,42
75	1,27
90	2,15
95	2,71
99	3,8
100	5,8

dung. Zudem berichtet CANCEIS über die Verteilung der Distanzen.

Bei den Gewichten und Merkmalen, die der Simulationsimputation zugrunde liegen, ergibt sich eine maximal mögliche Distanz von 106 für den Fall, dass alle Werte zwischen Empfänger und Spender sich (stark) unterscheiden. Sind alle Werte zwischen Empfänger und Spender identisch, ist die Distanz 0. Aus Tabelle 2 wird ersichtlich, dass das 1-%-Quantil der Verteilung der Distanz D_{fp} bei 0 liegt, also bei einem Prozent der Empfänger jeweils ein perfekter Spender mit Distanz 0 gefunden werden konnte. Eine Distanz ≤ 0,42 haben 50 % der Empfänger. Das ist insofern sehr gut, als dass bei diesen Fällen zwischen Empfänger und dem jeweiligen Spender maximal ein Merkmal abgewichen ist. In einem Prozent der Fälle liegt die Distanz über 3,8. Die maximal beobachtete Distanz bei einem Empfänger beträgt 5,8. Hier wurde also zwar wie geplant ein Spender aus demselben Betrieb gewählt, allerdings waren sich die anderen Merkmale fast nicht mehr ähnlich. Da diese Ergebnisse jedoch in sehr geringer Häufigkeit vorkommen, sind die Ähnlichkeiten zwischen den Empfängern und den gewählten Spendern insgesamt als sehr gut zu bewerten.

4.2 Bewertung anhand Fehlimputationsquote

In der Studie wurden jeweils 15 Imputationen für sechs Szenarien simulierter Ausfallwahrscheinlichkeiten von 10, 15, 20, 25, 30 und 50 % der gelieferten Arbeitnehmersätze je Betrieb erstellt. Für jeden Imputationslauf wurden dabei zufällig Werte unterschiedlicher Einheiten gelöscht. Die nachfolgende Ergebnisdarstellung erfolgt jeweils für die Durchschnittsergebnisse über die 15 Simulationsergebnisse im Szenario 20-%-Ausfallwahrscheinlichkeit. Zu erwähnen ist hierbei, dass sich die Ausfallwahrscheinlichkeit des gesamten Datensat-

zes durchaus von der ex ante je Betrieb gesetzten Ausfallwahrscheinlichkeit unterscheiden kann. Im Fall des ersten durchgeführten Laufs ergab sich beispielsweise für den Gesamtdatensatz eine Ausfallquote von 21,76 % (siehe Tabelle 2).

In ↘ Tabelle 3 sind die imputierten Eingliederungsnummern den tatsächlich gemeldeten gegenübergestellt. Von den insgesamt 292 093 Datensätzen waren 63 569 durch eine fehlende Eingliederungsnummer gekennzeichnet. Für alle konnte mit CANCEIS eine imputierte Eingliederungsnummer zugewiesen werden. Bei durchschnittlich 58 878 (92,6 %) stimmte die imputierte mit der gemeldeten Eingliederungsnummer überein. In 4 691 Fällen (7,4 %) entsprach die imputierte Eingliederungsnummer also nicht der tatsächlichen. Im Detail ergibt sich folgendes Bild:

- In 1 823 Fällen (2,9 %) wurde einem oder einer Beschäftigten, dessen oder deren Entlohnung in der Realität kein Tarifvertrag zugrunde lag, ein Tarifvertrag, der im Betrieb Anwendung findet, zugeordnet.
- In weiteren 1 113 Fällen (1,8 %) war es genau umgekehrt: Der beziehungsweise die Beschäftigte wurde nach einem Tarifvertrag entlohnt. Durch die Imputation wurde diese Person aber als nicht tarifvertraglich entlohnt ausgewiesen.
- In insgesamt 1 754 Fällen (2,8 %) wurde zwar korrekt imputiert, dass der oder die Beschäftigte nach Tarif entlohnt wurde, allerdings erfolgte hier die Zuspielung einer anderen Eingliederungsnummer aus dem Betrieb. Dieser Fehler kann in seiner Ergebniswirksamkeit noch relativiert werden, da es sich bei 758 Fällen der Fehlzuweisungen um Eingliederungsnummern „aus einer Familie“ handelt.[7]

↘ Tabelle 4 stellt den Anteil der Fehlimputationen in den unterschiedlichen Fehlerkategorien aus Tabelle 3 dar. Durchschnittlich kommt es bei rund 7 % der zu imputierenden Werte zu einer Fehlimputation. Zwischen den Imputationen in den Größenklassen ist hier kein systematischer Unterschied zu erkennen. Gleichwohl ist die nachfolgende systematische und ergebnisbeeinflussende Imputationsverzerrung

(gemeldet = kein VTV & Imputation = VTV) >
(gemeldet = VTV & Imputation = kein VTV)

zu beobachten. Diese Verzerrung begründet sich aus der Struktur des Datensatzes. Dieser besteht ausschließlich aus Betrieben, in denen Tarifverträge grundsätzlich Anwendung finden. Für diese Betriebe ist charakteristisch, dass die überwiegende Zahl der Arbeitnehmerinnen und Arbeitnehmer eine tarifliche Vergütung erhalten. Daher ist die Wahrscheinlichkeit höher, dass CANCEIS

7 Es könnte sich beispielsweise um einen Lohntarifvertrag (als Vergütungstarifvertrag) und einen Gehaltstarifvertrag (als einen zweiten Vergütungstarifvertrag) handeln, die aber beide in einem gemeinsamen Tarifvertrag (TV) verankert sind.

Tabelle 3

Ergebnisse der Imputationsstudie nach Größenklassen bei gesetzter Ausfallquote von 20 %
Durchschnitt über 15 Simulationsläufe

	Insgesamt	Fehlende Eingliederungsnummer	Imputation							
			korrekt	fehlerhaft	darunter:					
					aus anderem Betrieb	aus gleichem Betrieb				
						gemeldet:	kein VTV	VTV	VTV	darunter:
						Imputation:	VTV	kein VTV	VTV	anderer VTV, aber gleicher TV
1 – 9 Beschäftigte	6 734	1 903	1 736	167	–	167	85	56	26	9
10 – 49 Beschäftigte	38 852	9 179	8 576	603	–	603	270	153	180	85
50 – 99 Beschäftigte	34 890	7 624	7 151	473	–	473	210	116	147	75
100 – 249 Beschäftigte	58 362	12 666	11 781	885	–	885	368	218	299	155
250 – 499 Beschäftigte	46 010	9 780	9 065	715	–	715	280	167	268	115
500 – 999 Beschäftigte	30 774	6 548	6 047	501	–	501	183	115	203	88
1 000 Beschäftigte und mehr	76 471	15 869	14 522	1 347	–	1 347	428	289	631	232
Insgesamt	292 093	63 569	58 878	4 691	–	4 691	1 823	1 113	1 754	758

VTV: Vergütungstarifvertrag; TV: Tarifvertrag

Tabelle 4

Anteil der Fehlimputationen der Studie nach Größenklassen bei gesetzter Ausfallquote von 20 %
Durchschnitt über 15 Simulationsläufe

	Imputationserfolgs-/Imputationsfehlerquote bezogen auf zu imputierende Datensätze							
	korrekt	fehlerhaft	darunter:					
			aus anderem Betrieb	aus gleichem Betrieb				
				gemeldet:	kein VTV	VTV	VTV	darunter:
				Imputation:	VTV	kein VTV	VTV	anderer VTV, aber gleicher TV
	%							
1 – 9 Beschäftigte	91,2	8,8	–	8,8	4,5	2,9	1,4	0,5
10 – 49 Beschäftigte	93,4	6,6	–	6,6	2,9	1,7	2,0	0,9
50 – 99 Beschäftigte	93,8	6,2	–	6,2	2,8	1,5	1,9	1,0
100 – 249 Beschäftigte	93,0	7,0	–	7,0	2,9	1,7	2,4	1,2
250 – 499 Beschäftigte	92,7	7,3	–	7,3	2,9	1,7	2,7	1,2
500 – 999 Beschäftigte	92,3	7,7	–	7,7	2,8	1,8	3,1	1,3
1 000 Beschäftigte und mehr	91,5	8,5	–	8,5	2,7	1,8	4,0	1,5
Insgesamt	92,6	7,4	–	7,4	2,9	1,8	2,8	1,2

VTV: Vergütungstarifvertrag; TV: Tarifvertrag

einen geeigneten Spender mit tariflicher Entlohnung (bei sonst ähnlichen Merkmalen) wählt. Diese Verzerrung ist über alle Simulationsszenarien, Simulationsläufe und auch Größenklassen hinweg zu beobachten. Daraus kann jedoch noch keine Aussage zur Ergebniswirksamkeit dieser Verzerrung abgeleitet werden. Dies müsste in weiteren Analysen untersucht werden (siehe Kapitel 5).

Grundsätzlich können für eine relative Bewertung des Fehlimputationseffektes die Fehlimputationen nicht nur auf die insgesamt zu imputierenden Datensätze (siehe Tabelle 4) bezogen werden, sondern auch auf die Gesamtzahl aller Datensätze. Die nachfolgende Tabelle 5 stellt beide Fehlimputationsquoten jeweils insgesamt und bereinigt um den „Tarifvertragsfamilien-Effekt“ für das durchschnittliche 20-%-Ausfallszenario dar.

Während bei der Fehlerquote bezogen auf die zu imputierenden Werte eine Aussage zur Treffsicherheit des Imputationsverfahrens als solches abgeleitet wird und diese im Untersuchungsfall bei

Treffsicherheit des Imputationsverfahrens =
100 % – 6,2 % = 93,8 %

über alle Größenklassen liegt, zeigt die zweite Betrachtungsweise, wie hoch der Schaden einer Fehlimputation für das Gesamtergebnis (der ETI) einzuschätzen ist. Über alle Größenklassen hinweg ergibt sich eine maximale (bereinigte) Ergebnisverzerrung von 1,3 %.

Dies ist als eine Obergrenze für die ETI zu bewerten, da im Analysebeispiel angenommen wird, dass in jedem Betrieb rund 20 % der Datensätze eine fehlende Eingliederungsnummer aufweisen. Für die Erhebung der ETI bedeutet eine Anwendung der 20-%-Quote jedoch nur, dass jede Betriebsmeldung abgewiesen und nochmals beim Melder anzufordern ist, wenn er mehr als 20 % Ausfallquote aufweist. Es ist allerdings nicht anzunehmen, dass alle (von den Statistischen Ämtern der Länder) akzeptierten Meldungen bei 20 % der Arbeitnehmersätze fehlende Informationen zum angewandten Tarifvertrag aufweisen.

In ↘ Tabelle 5 ist jedoch noch ein weiterer Effekt zu beobachten: Die Fehlimputationsquote bei den Kleinstbetrieben (bis unter 10 Beschäftigte) ist im Vergleich zu den anderen Größenklassen geringfügig höher. Gleichwohl lassen (leicht verzerrte) Einzelmeldungen dieser Kleinstbetriebe einen relativ geringen Einfluss auf das gesamtstatistische Ergebnis vermuten (siehe hierzu Kapitel 5).

Tabelle 5

Fehlimputationsquote nach Größenklassen bei gesetzter Ausfallquote von 20%
Durchschnitt über 15 Simulationsläufe

	Bezogen auf Imputationen		Bezogen auf Gesamtdatenbestand	
	insgesamt	bereinigt gleicher Tarifvertrag	insgesamt	bereinigt gleicher Tarifvertrag
	%			
1 – 9 Beschäftigte	8,8	8,3	2,5	2,3
10 – 49 Beschäftigte	6,6	5,6	1,6	1,3
50 – 99 Beschäftigte	6,2	5,2	1,4	1,1
100 – 249 Beschäftigte	7,0	5,8	1,5	1,3
250 – 499 Beschäftigte	7,3	6,1	1,6	1,3
500 – 999 Beschäftigte	7,7	6,3	1,6	1,3
1 000 Beschäftigte und mehr	8,5	7,0	1,8	1,5
Insgesamt	7,4	6,2	1,6	1,3

↘ Tabelle 6 stellt schließlich die Ergebnisse der Imputation in den unterschiedlichen Simulationsszenarien zusammenfassend dar. Hierbei wurden die arithmetischen Mittel der einzelnen Imputationsläufe herangezogen. Zur Beurteilung der Stabilität der Simulationsergebnisse sind in ↘ Tabelle 7 Varianz und Standardabweichung beigefügt.

Entsprechend der Spezifikationsvorgaben des CANCEIS-Laufes ist keine (fehlerhafte) Imputation aus einem fremden Betrieb zu beobachten. Es konnte zudem für jeden fehlenden Wert ein Spender gefunden werden, auch im Szenario mit 50-prozentiger Ausfallwahrscheinlichkeit.

Die bereinigte Fehlimputationsquote bezogen auf die zu imputierenden Fälle ist bei allen betrachteten Ausfallquoten nahezu identisch und liegt durchschnittlich zwischen 6,0 und 7,0 %. ↘ Tabelle 8 Das heißt, der Imputationsprozess wurde nicht systematisch durch zu wenig vorhandene Spender negativ beeinflusst. Folgerichtig steigt jedoch die Quote der Fehlimputationen bezogen auf den Gesamtdatensatz mit steigender Ausfallquote,

Tabelle 6

Durchschnittliche Fehlimputationsquoten bei unterschiedlichen Ausfallquoten-Szenarien

Szenario	Insgesamt	Fehlende Eingliederungsnummer	Imputation							
			korrekt	fehlerhaft	darunter:					
					aus anderem Betrieb	aus gleichem Betrieb				
						wahr: Imputation:	kein VTV VTV	VTV kein VTV	VTV VTV	darunter: anderer VTV, aber gleicher TV
	Anzahl der Arbeitnehmerdatensätze									
50 Prozent	292 093	149 308	136 797	12 511	–	12 511	4 853	2 908	4 750	2 051
30 Prozent	292 093	93 578	86 428	7 150	–	7 150	2 784	1 704	2 662	1 154
25 Prozent	292 093	77 971	72 149	5 822	–	5 822	2 274	1 371	2 177	951
20 Prozent	292 093	63 569	58 878	4 691	–	4 491	1 823	1 113	1 754	758
15 Prozent	292 093	50 905	47 191	3 714	–	3 714	1 432	894	1 387	605
10 Prozent	292 093	34 893	32 382	2 511	–	2 511	972	615	924	403
	in Prozent aller Arbeitnehmerdatensätze									
50 Prozent	100	51,1	46,8	4,3	–	4,3	1,7	1,0	1,6	0,7
30 Prozent	100	32,0	29,6	2,4	–	2,4	1,0	0,6	0,9	0,4
25 Prozent	100	26,7	24,7	2,0	–	2,0	0,8	0,5	0,7	0,3
20 Prozent	100	21,8	20,2	1,6	–	1,6	0,6	0,4	0,6	0,3
15 Prozent	100	17,4	16,2	1,3	–	1,3	0,5	0,3	0,5	0,2
10 Prozent	100	11,9	11,1	0,9	–	0,9	0,3	0,2	0,3	0,1

VTV: Vergütungstarifvertrag; TV: Tarifvertrag

Tabelle 7
Statistische Stabilitätsparameter der Simulation

Simulationsszenario	Statistische Kennzahl	Fehlimputationsquote				Verzerrungsanteil
		bezogen auf Imputationen		bezogen auf Datensatz		
		insgesamt	bereinigt gleicher Tarifvertrag	insgesamt	bereinigt gleicher Tarifvertrag	
		%				
Ausfallquote = 10 Prozent	Mittelwert	7,19	6,04	0,86	0,72	0,12
	Varianz	0,00	0,00	0,00	0,00	0,00
	Standardabweichung	0,07	0,09	0,01	0,01	0,02
Ausfallquote = 15 Prozent	Mittelwert	7,30	6,11	1,27	1,06	0,18
	Varianz	0,00	0,00	0,00	0,00	0,00
	Standardabweichung	0,07	0,07	0,09	0,07	0,02
Ausfallquote = 20 Prozent	Mittelwert	7,38	6,19	1,61	1,35	0,24
	Varianz	0,00	0,00	0,00	0,00	0,00
	Standardabweichung	0,11	0,09	0,02	0,02	0,02
Ausfallquote = 25 Prozent	Mittelwert	7,47	6,25	1,99	1,67	0,31
	Varianz	0,00	0,00	0,00	0,00	0,00
	Standardabweichung	0,08	0,08	0,02	0,02	0,02
Ausfallquote = 30 Prozent	Mittelwert	7,64	6,41	2,45	2,05	0,37
	Varianz	0,00	0,00	0,00	0,00	0,00
	Standardabweichung	0,08	0,07	0,03	0,02	0,04
Ausfallquote = 50 Prozent	Mittelwert	8,38	7,01	4,28	3,58	0,67
	Varianz	0,00	0,00	0,00	0,00	0,00
	Standardabweichung	0,05	0,05	0,02	0,02	0,03

Tabelle 8
Durchschnittliche Fehlimputationsquoten sowie Anteil der Imputationsverzerrung

Simulations-szenario	Bezogen auf Imputa-tionen		Bezogen auf Datensatz		Verzer-rungs-anteil
	insgesamt	bereinigt gleicher Tarifver-trag	insgesamt	bereinigt gleicher Tarifver-trag	
	%				
50 Prozent	8,4	7,0	4,3	3,6	0,7
30 Prozent	7,6	6,4	2,4	2,1	0,4
25 Prozent	7,5	6,2	2,0	1,7	0,3
20 Prozent	7,4	6,2	1,6	1,3	0,2
15 Prozent	7,3	6,1	1,3	1,1	0,2
10 Prozent	7,2	6,0	0,9	0,7	0,1

beträgt aber selbst im Fall einer 50-prozentigen Ausfallquote nur 3,6 %. Ergänzend ist in den Tabellen 7 und 8 der Verzerrungsanteil als Indikator zum Ausweis der oben benannten Verzerrung nach der untenstehenden Formel ausgewiesen.

Wie erwartet, steigt der Anteil der Verzerrung hin zu einem höheren Anteil an Vergütungstarifverträgen in den Szenarien mit höheren Ausfallquoten.

$$\text{Verzerrungsanteil} = \frac{\text{"gemeldet = kein VTV und Imputation = VTV"} - \text{"gemeldet = VTV und Imputation = kein VTV "}}{\text{Gesamtzahl der Datensätze}}$$

5

Zusammenfassung und Ausblick

Aus der besonderen Erhebungssituation der ETI, die die nachträglich erhobenen Angaben zu Tarifinformationen mit den Ergebnissen der Verdiensterhebung kombinieren muss, ergibt sich die erhebungstechnisch bedeutsame Frage der Vollständigkeit des zusammengeführten Betriebsdatensatzes.

Die vorliegende datenbasierte Studie untersucht, ob der Einsatz von Imputationen ein probates Mittel zur Kompensation von Antwortausfällen für die Erhebung der Tarifinformationen 2025 darstellen könnte. Die bisherigen Ergebnisse anhand der Bewertung der Imputationserfolgs- beziehungsweise Imputationsfehlerquoten liefern keine Anzeichen, dass Imputationen nicht geeignet sind.

Die Untersuchung hat gezeigt, dass es selbst bei einer Ausfallquote von 50 % möglich war, durch die Imputation einen Spender aus dem Betriebsdatensatz zu finden. Auch mit zunehmender Ausfallquote ist die (bereinigte) Fehlimputationsquote bezogen auf alle zu imputierenden Datensätze relativ stabil. Dies weist darauf hin, dass die Imputationsqualität nicht durch die abnehmende Zahl der potenziellen Spender stark einbricht. Als drittes Ergebnis bleibt festzuhalten, dass selbst bei der kleinsten betrieblichen Größenklasse (1 bis 9 sozialversicherungspflichtig Beschäftigte) die Imputation grundsätzlich funktioniert.

Gleichwohl kann aus diesen Untersuchungsergebnissen noch nicht abgeleitet werden, ob und bis zu welcher Ausfallquote die Imputation im Rahmen der ETI eingesetzt werden kann. Vielmehr ist in einem nächsten Schritt zu untersuchen, welche Auswirkungen die Fehlimputationen auf die (zentralen) statistischen Ergebnisse haben. Hierbei sind insbesondere mögliche Verzerrungseffekte der Imputation in drei Dimensionen zu analysieren:

1. Auswirkungen auf den Anteil der Beschäftigten mit Tarifvertrag (im Betrieb, im Wirtschaftsabschnitt, im Bundesland, nach Größenklassen beziehungsweise in der Gesamtwirtschaft),
2. Auswirkungen auf zentrale Verdienstkenngrößen für die Gruppe der Beschäftigungsverhältnisse mit und ohne tarifliche Entlohnung. Zu untersuchen wären hierbei Auswirkungen auf
   - den durchschnittlichen Bruttostundenverdienst beziehungsweise auch Verteilungen in den beiden Teilgruppen differenziert nach Wirtschaftsabschnitt, Bundesland und Größenklasse,
   - den Anteil der Beschäftigten unterhalb der Niedriglohngrenze beziehungsweise mit Mindestlohn differenziert nach Wirtschaftsabschnitt, Bundesland und Größenklasse,
   - die Altersstruktur (und andere individuelle Merkmale der Beschäftigten) in den beiden Teilsegmenten (mit und ohne Tarifvertrag) differenziert nach Wirtschaftsabschnitt, Bundesland und Größenklasse,
3. implizite Effekte auf den Tarifindex durch ein imputationsverändertes Wägungsschema.

Während bei den ersten beiden Punkten nur entscheidend ist, ob eine tarifliche Entlohnung vorliegt oder nicht, ist bei Punkt 3 auch entscheidend, welcher Tarifvertrag dem Arbeitnehmersatz zugeordnet ist. Es ist zu untersuchen, ob durch potenzielle Fehlimputation

- Tarifverträge unter Umständen nicht mehr ins Wägungsschema des Tarifindex aufgenommen werden würden,
- sich die Gewichte einzelner Tarifverträge durch die mögliche Verzerrung der Imputation maßgeblich verschieben könnten, und letztlich
- ist abzuschätzen, wie stark sich ein geändertes Wägungsschema auf die Entwicklung des Tarifindex auswirken könnte.

Erst im Anschluss der Bewertung der Ergebnisse dieser Untersuchungen kann eine begründete Entscheidung über den Einsatz des Imputationsverfahrens bei unvollständigen Betriebsmeldungen im Rahmen der ETI getroffen werden. Wie hoch sich die Entlastung der Wirtschaft und die der Statistischen Ämter der Länder durch den Einsatz der Imputation anstelle einer Neuanforderung der Meldung beziffert, kann erst in einer Ex-post-Analyse nach Durchführung der Erhebung zu Tarifinformationen analysiert werden.

LITERATURVERZEICHNIS

Bankier, Michael/Poirier, Paul/Lachance, Martin. *Efficient Methodology within the Canadian Census Edit and Imputation System (CANCEIS)*. 2001. Proceedings of the Annual Meeting of the American Statistical Association. [Zugriff am 5. Dezember 2023]. Verfügbar unter: www.asasrms.org

CANCEIS Development Team. *CANCEIS User's Guide*. Version 5.2. Ottawa 2015.

Finke, Claudia/Geisler, Susanna/Überschaer, Anja. *Aus drei mach eins: die neue Verdiensterhebung*. In: WISTA Wirtschaft und Statistik. Ausgabe 5/2023, Seite 58 ff.

Frentzen, Kathrin/Günther, Roland. *Korrektur des Antwortausfalls in der Verdiensterhebung 2015*. In: WISTA Wirtschaft und Statistik. Ausgabe 2/2017, Seite 24 ff.

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

Preising, Marcel/Lange, Kerstin/Dumpert, Florian. *Imputation zur maschinellen Behandlung fehlender und unplausibler Werte in der amtlichen Statistik*. In: WISTA Wirtschaft und Statistik. Ausgabe 5/2021, Seite 40 ff.

Schymura, Sandra. *Beschäftigte und ihre Verdienste nach der zweiten Erhöhung des Mindestlohns*. In: WISTA Wirtschaft und Statistik. Ausgabe 6/2020, Seite 58 ff.

Statistisches Bundesamt. *Klassifikation der Wirtschaftszweige 2008*. Wiesbaden 2008. [Zugriff am 18. Dezember 2023]. Verfügbar unter: www.destatis.de

Statistisches Bundesamt. *Qualitätsbericht Verdienststrukturerhebung*. 2018. [Zugriff am 29. November 2023]. Verfügbar unter: www.destatis.de

Statistisches Bundesamt. *Index der Tarifverdienste – Methodische Erläuterungen*. 2021. [Zugriff am 5. Dezember 2023]. Verfügbar unter: www.destatis.de

RECHTSGRUNDLAGEN

Gesetz über die Statistik der Verdienste und Arbeitskosten (Verdienststatistikgesetz – VerdStatG) vom 21. Dezember 2006 (BGBl. I Seite 3291), das zuletzt durch Artikel 1 des Gesetzes vom 12. August 2020 (BGBl. I Seite 1872) geändert worden ist.

Gesetz zur Regelung eines allgemeinen Mindestlohns (Mindestlohngesetz – MiLoG) vom 11. August 2014 (BGBl. I Seite 1348), das zuletzt durch Artikel 2 des Gesetzes vom 28. Juni 2023 (BGBl. I Nr. 172) geändert worden ist.

Herausgeber
Statistisches Bundesamt (Destatis), Wiesbaden

Schriftleitung
Dr. Daniel Vorgrimler
Redaktion: Ellen Römer

Ihr Kontakt zu uns
www.destatis.de/kontakt

Erscheinungsfolge
zweimonatlich, erschienen im Februar 2024
Ältere Ausgaben finden Sie unter www.destatis.de sowie in der Statistischen Bibliothek.

Artikelnummer: 1010200-24001-4, ISSN 1619-2907