



Jörg Hagenah, Denise Sommer, Lea Hildebrand

Das theoriegeleitete Prüfverfahren zur Messung der Qualität von Datenfusionen

Ein Replikations-Test am Beispiel „Schützt politisches Wissen vor Populismus?“ (Westle 2020)



Ostfalia
Hochschule für angewandte
Wissenschaften

Jörg Hagenah, Denise Sommer, Lea Hildebrand

Das theoriegeleitete Prüfverfahren zur Messung der Qualität von Datenfusionen

Ein Replikations-Test am Beispiel der Studie „Schützt politisches Wissen vor Populismus?“ (Westle 2020)



Jörg Hagenah, Denise Sommer, Lea Hildebrand

Das theoriegeleitete Prüfverfahren zur Messung der Qualität von Datenfusionen. Ein Replikations-Test am Beispiel der Studie „Schützt politisches Wissen vor Populismus?“ (Westle 2020)

Working Paper (Fachbeiträge Verkehr-Sport-Tourismus-Medien)

Nr. 1 (09/2023)

Herausgeberin:

Fakultät Verkehr-Sport-Tourismus-Medien, Ostfalia Hochschule für angewandte Wissenschaften, Karl-Scharfenberg-Str. 55/57, 38229 Salzgitter, <https://www.ostfalia.de/cms/de>

Autor/ Autoren:

Jörg Hagenah, Dr., Lehrkraft für besondere Aufgaben an der Fakultät K, Methoden der empirischen Sozialforschung/Medien- und Kommunikationswissenschaft; Mediaforschung; Sozialer und medialer Wandel; Politische Kommunikation, Medien und Sport, j.hagenah@ostfalia.de

Denise Sommer, Dr., Professorin für Theorie der Kommunikations- und Medienwissenschaft, Medienwirkung und Mediennutzung im Wandel, Medien und Diversity, interpersonale Kommunikation, Forschungsmethoden, de.sommer@ostfalia.de

Lea Hildebrand, M.A., Wissenschaftliche Mitarbeiterin an der Fakultät K und Doktorandin der TU Clausthal, Wirtschaftlichkeitsanalysen im Public und Non-Profit Management, Performance Management von Museen, Effizienz- und Zeitreihenmodellierungen, le.hildebrand@ostfalia.de

Lizenz

© 2023 beim Autor

Bildnachweis Titel: Frauenhofer-Zentrum MEOS (2023). Multimodale Datenfusion und Analyse

Zitiervorschlag:

Hagenah, J., Sommer, D. & Hildebrand, L. (2023). Das theoriegeleitete Prüfverfahren zur Messung der Qualität von Datenfusionen. Ein Replikations-Test am Beispiel der Studie „Schützt politisches Wissen vor Populismus?“ (Westle 2020). Working Paper (Fachbeiträge Verkehr-Sport-Tourismus-Medien), Nr. 1 (09/2023)

Tabellenverzeichnis

Tabelle 1: Regression des politischen Vorwissens (uV) auf den Populismus (aV) im Vergleich zwischen Originaldatensatz und drei fusionierten Datensätzen3

Tabelle 2: Regression von unabhängigen Variablen (uV) auf den Populismus (aV): Berechnung des pore-Scores für b-Koeffizienten im Vergleich zwischen Originaldatensatz und drei fusionierten Datensätzen6

Tabelle 3: Regression von unabhängigen Variablen (uV) auf den Populismus (aV): pore-Scores für b-Koeffizienten, r²-Werten und signifikanz im Vergleich zwischen Originaldatensatz und drei fusionierten Datensätzen7

Das theoriegeleitete Prüfverfahren zur Messung der Qualität von Datenfusionen.

Ein Replikations-Test am Beispiel der Studie „Schützt politisches Wissen vor Populismus?“ (Westle 2020)

Von Jörg Hagenah¹, Denise Sommer & Lea Hildebrand

Ostfalia Hochschule für angewandte Wissenschaften, Working Paper

In dieser Vorstudie wird in den theoriegeleiteten Untersuchungsansatz zur Prüfung von Datenfusionen eingeführt, um replikationsorientierte Qualitätskriterien zu erarbeiten. Bei dem theoriegeleiteten Ansatz handelt es sich nicht um primäre Theoriearbeit, sondern um einen Weg, diese notwendige Theoriearbeit auszulagern. Es wird auf theoriegeleitete Hypothesen zurückgegriffen, die mit repräsentativen Datenerhebungen in bereits publizierten Artikeln geprüft wurden. Aus theoretischer Perspektive lässt sich hierdurch argumentieren, dass die zu replizierenden Hypothesentests kausale Mechanismen zeigen, die vorab oder ex-post auf der Basis von theoretisch abgeleiteten Hypothesen aufgestellt oder interpretiert wurden. Das Prinzip, bei der Theoriearbeit auf publizierte Studien zurückzugreifen, soll die Transparenz des Forschungsprozesses erhöhen. Ähnlich wie bei dem in der Psychologie mittlerweile üblichem Verfahren der Präregistrierung von Forschungsdesigns (zpid 2023), soll der potenzielle Verdacht vermieden werden, dass Hypothesen schon mit Blick auf vorliegende Ergebnisse abgeleitet sein könnten (Kerr 1998).

Im Mittelpunkt des vorliegenden Beitrags steht die Einführung eines Prüfverfahrens zur Qualitätsmessung von Datenfusionen. Dafür erfolgt zunächst eine kurze methodische Vorstellung der Datenfusionsverfahren, die für unseren Prüfzweck verwendet wurden. Anschließend wird am Beispiel der Studie „Schützt politisches Wissen vor Populismus?“ von Westle (2020) eine Replikation der Hypothesenprüfung mit verschiedenen fusionierten Datensätzen vorgenommen, um 1.) die Brauchbarkeit fusionierter Datensätze für komplexere Kausalanalysen zu testen und 2.) verschiedene Verfahren der Datenfusion miteinander hinsichtlich ihrer Tauglichkeit zu vergleichen.

Idee und Verfahrensweisen von Datenfusionen

Datenfusionen zielen darauf ab, durch die Kombination mehrerer Datenquellen vollständige Datensätze zu gewinnen, was primär auf der Lösung eines spezifischen Imputationsproblems beruht (Bacher & Prander 2018). Dies lässt sich damit begründen, dass durch die Fusion von multisensorischen Daten zunächst ein neuer (kombinierter) Datensatz entsteht, der fehlende Werte für die im Fusionsprozess festgelegten spezifischen Variablen enthält. Unter *spezifischen* oder *geteilten Variablen* sind in diesem Kontext jene Variablen zu verstehen, welche die zusammenzuführenden Datenquellen *teilen*. Der kombinierte Datensatz beinhaltet somit ausschließlich vollständige Informationen zu allen gemeinsamen Variablen, die zur Durchführung der Imputation verwendet werden.

In den vergangenen Jahren sind verschiedene methodische Ansätze entstanden, um solche vollständigen multisensorischen Datensätze in vielfältigen Studienkontexten zu generieren (siehe z.B. Khaleghi 2013). Für unsere Studie haben wir uns entschieden, drei etablierte Imputationsverfahren zu berücksichtigen. Da verschiedene Open-Source-Softwarepakete für die technische Umsetzung der genutzten Methoden verwendet werden können, ist der empirische Ansatz der vorliegenden Studie einfach zu replizieren. Im Rahmen der vorliegenden Untersuchung

¹ Kontakt: j.hagenah@ostfalia.de

wurden alle Imputationen in R durchgeführt. Zur technischen Umsetzung der Imputationen wurden die entsprechenden Funktionen des Pakets *simputation* (Van der Loo 2022) verwendet. Der folgende Abschnitt enthält eine zusammenfassende Beschreibung der angewandten Methoden. Allen gemeinsam ist, dass die imputierten Werte auf mindestens eine ähnliche Beobachtung bezogen werden, die vollständige Informationen für die jeweilige spezifische Variable enthält.

(1) *Sequentielle Hot-Deck-Imputation (SHD)*

SHD ist eine beliebte Imputationsmethode aufgrund ihrer Anwendbarkeit auf verschiedene Arten von Daten (Rhadatunnisa & Wilantika 2022). So ermöglicht z.B. das R-Paket *simputation* die Verarbeitung numerischer, kategorialer und gemischter Daten (Van der Loo 2022). Fehlende Werte werden sukzessive durch Werte ähnlicher Beobachtungen ersetzt. Zunächst werden die Daten unter Berücksichtigung des festgelegten Satzes gemeinsamer Variablen sortiert. Anschließend werden die fehlenden Werte durch die Werte der vorherigen Beobachtung im Datensatz ersetzt (Andrige & Little 2010).

(2) *Predictive Mean Matching (PMM)*

PMM wurde ursprünglich von Little (1988) eingeführt und ist eine besonders verbreitete Imputationsmethode. PMM basiert auf der anfänglichen Erstellung eines prädiktiven (Regressions-)Modells. Fehlende Werte werden zunächst durch vorhergesagte Modellwerte ersetzt und schließlich durch eine Kombination von beobachteten, nächstgelegenen Werten mit den zuvor erhobenen Vorhersagen imputiert (Jadhav 2019). In Bezug auf die für diese Studie verwendete Funktion des Pakets *simputation* umfasst der Pool der nächstgelegenen Werte Beobachtungen, die die geringste absolute Abweichung von der jeweiligen Vorhersage aufweisen (Van der Loo 2022).

(3) *K-Nächster Nachbar (KNN)*

KNN basiert auf der Suche nach den k nächsten Nachbarn innerhalb des zugrundeliegenden Datensatzes (wobei k eine von den Forschenden gewählte Zahl ist). Im *simputation* Paket wird der Ähnlichkeitskoeffizient von Gower (1971) verwendet, um den erforderlichen Pool von den k nächsten Nachbarn zu bestimmen (Van der Loo 2022). Der schließlich imputierte Wert ist als eine Kombination (z. B. der Mittelwert) der Werte der nächsten Nachbarn zu verstehen (Zhang 2012).

Anwendung der Datenfusion für eine Replikationsstudie

Um die Idee eines theoriegeleiteten Prüfverfahrens für Datenfusionen zeigen zu können, wurden Teile der Studie „Schützt politisches Wissen vor Populismus?“ von Westle (2020) repliziert. Die folgenden Kriterien haben zur Auswahl dieses Artikels geführt:

- Ein wichtiger Grund ist die verwendete Datengrundlage: die Allgemeine Bevölkerungsumfrage der Sozialwissenschaften, ALLBUS 2018 (Westle 2020: 212). Sie zählt als eine etablierte Datenquelle, die eine besonders große Datenqualität hat, leicht zugänglich ist und häufig für Sekundäranalysen verwendet wird (Diekmann et al. 2019).
- Der Artikel wurde publiziert in einem Herausgeberband mit einem double-blind Peer-Review-Verfahren, für das (neben den Herausgebern) auch externe Gutachter eingesetzt wurden. Aus theoretischer Perspektive erfolgt somit eine doppelte Belastbarkeit der Hypothesen, deren theoretische Ableitung durch die externe Begutachtung geprüft wurde.

- In dem Artikel werden eine Vielzahl von Hypothesen, v.a. auf Basis der zentralen Konstrukts *Populismus* abgeleitet, so dass eine breite Basis an auszahlbaren Hypothesentests vorhanden ist.
- Bei der zentralen unabhängigen Variable (uV) *politisches Vorwissen* (Moosdorf at al. 2020) und der zentralen abhängigen Variablen (aV) *Populismus* (Westle 2020: 212-216) handelt es sich darüber hinaus um *multiple-item-indices*. Sie ermöglichen eine stabilere Analyse als sog. *single-item-Variablen* an zentraler Stelle.
- Die Beschreibung und Aufbereitung der genutzten Variablen ist anhand von Prozentwerten, Mittelwerten, Standardabweichungen und anhand von N fehlend und N gültig transparent dokumentiert.
- Alle Einzelhypothesen wurden anhand von bivariaten Regressionen sowie gebündelt in multiplen Regressionen berechnet.

Tabelle 1 zeigt die Prüfung der Haupthypothese und das Untersuchungsdesign. Um den richtigen Analyseweg im Blick zu haben, wurde die bivariate Regression mit Hilfe des Datensatzes Allbus 2018 nachgerechnet. In der Tabelle stehen in den Zeilen das Signifikanzniveau (Anzahl Sterne), der b-Koeffizient und das r^2 . Zur Vorbereitung der Datenfusionen wurde aus dem Allbus 2018 eine Zufallsstichprobe gezogen (*Allbus Hälfte 1*) - alle nicht gezogenen Fälle gehören zu *Hälfte 2*. Als Referenzgrößen können die Regressionsergebnisse bei den beiden Hälften fungieren. Anders als bei dem Gesamtdatensatz lassen sich kleinere Abweichungen (um max. 0,01-Punkte) im Vergleich zu dem publizierten b-Koeffizient und zu dem r^2 finden.

Der Hauptblick liegt aber auf den neu fusionierten Datensätzen mittels K-Nearest Neighbor (KNN), Predictive Mean Matching (PMM) und Sequential Hot-Deck Imputation (SHD), deren Ergebnisse etwas weniger (SHD) oder etwas mehr (KNN, PMM) von den Originalergebnissen abweichen als bei den beiden Zufallshälften. Im Kern würden die Ergebnisse aber zu derselben Interpretation führen. Politisches Vorwissen schützt etwas vor Populismus: Es gibt einen kleinen negativen Zusammenhang, der etwas weniger als 10 % der Varianz aufklärt und der höchst signifikant ist ($p < 0,001$).

Tabelle 1: Regression des politischen Vorwissens (uV) auf den Populismus (aV) im Vergleich zwischen Originaldatensatz und drei fusionierten Datensätzen

| gewichtet | Westle 2020: 226 | Replikation mit Allbus 2018 | Differenz Replikation zu Westle 2020 | Allbus 2018, Hälfte 1 | Diff. Hälfte 1 zu Westle 2020 | Allbus 2018, Hälfte 2 | Diff. Hälfte 2 zu Westle 2020 | Diff. KNN zu Westle 2020 | Diff. PMM zu Westle 2020 | Diff. SHD zu Westle 2020 | | | |
|---------------|------------------|-----------------------------|--------------------------------------|-----------------------|-------------------------------|-----------------------|-------------------------------|--------------------------|--------------------------|--------------------------|-------|------|------|
| Anzahl Sterne | 3 | 3 | 0,00 | 3 | 0,00 | 3 | 0,00 | 3 | 0,00 | 3 | 0 | 3 | 0,00 |
| b-Koeffizient | -0,22 | -0,22 | 0,00 | -0,21 | -0,01 | -0,23 | 0,01 | -0,20 | -0,02 | -0,25 | 0,03 | -0,2 | 0,01 |
| r^2 | 0,08 | 0,08 | 0,00 | 0,07 | 0,01 | 0,09 | -0,01 | 0,06 | 0,02 | 0,11 | -0,03 | 0,08 | 0,00 |

uV = unabhängige Variable; aV = abhängige Variablen
Beide Indizes wurden in kontinuierliche 0 bis 1 Variablen umgewandelt

Anmerkungen: KNN = K-Nearest Neighbor; PMM = Predictive Mean Matching; SHD = Sequential Hot Deck Imputation

Aufgrund des Forschungsdesigns erleichtern diese Bedingungen die Datenfusionen. Es gibt keine Probleme mit einer differierenden Grundgesamtheit, auch Datenerhebungsverfahren und Datenqualität sind identisch. Außerdem stünden eine Vielzahl an gemeinsamen Variablen zur Verfügung. Auf diese Optimierungsmöglichkeit wird aufgrund einer kleiner werdenden

Verallgemeinerbarkeit bei einem Heranziehen einer großen Anzahl an gemeinsamen Variablen verzichtet. Ziel ist es, mit einer begrenzten Anzahl an gemeinsamen Variablen ein möglichst großes Anwendungspotenzial derartiger Verfahren im Blick zu haben.

Ideal wäre es nur diejenigen Variablen als gemeinsame Variablen heranzuziehen, die in harmonisierter Form in allen großen sozialwissenschaftlichen Studien vorhanden sind. Wir verstoßen damit bewusst gegen Teile häufig genannter Voraussetzungen (hohe Interkorrelationen; große inhaltliche Breite; Axenfeld et al. 2022: 1245) und probieren einen simplen Weg aus. Dadurch sinkt zwar die Wahrscheinlichkeit einer optimalen Fusion. Allerdings steigt dafür die Anzahl an Fusionsmöglichkeiten, wenn wir uns auf diejenigen Variablen konzentrieren, die in (allen) größeren Datensätzen vorhanden sind. Dafür gilt es, diese Variablen in einem ersten Schritt herauszufiltern. Für die vorliegende Analyse haben wir den Weg (vorläufig) folgendermaßen abgekürzt: Ein Vergleich zwischen dem Allbus 2018 und dem aktuellsten (aufbereiteten) Datensatz der Media-Analyse 2015 (https://search.gesis.org/research_data/ZA5761; vgl. Hagenah & Meulemann 2009) zeigt, dass 18 gemeinsame Variablen aus den folgenden Variablenbereichen zu finden sind:

- Informationen über Gemeinden und Gebiete (3 Variablen),
- Soziodemographie (10),
- Haushalt (2) und
- Einstellungen und Verhalten (3).

Davon wurden nur diejenigen neun Variablen verwendet, die bei Westle (2020) als unabhängige Variablen oder Kontrollvariablen verwendet wurden und somit publizierte Vergleichsergebnisse auch für das Hypothesentesten bieten. Da Westle alle Variablen in eine 0/1-Form gebracht hat - entweder in einer kontinuierlichen Form bei metrischen Variablen oder als Dummyvariablen bei kategorialen (ordinalen oder nominalen) Variablen - ist keine weitere Standardisierung notwendig und alle Variablen werden gleichgewichtig in die Berechnung von nötigen Koeffizienten einbezogen. Als spezifische Variablen wurden die zentrale uV (politisches Vorwissen) und die zentrale aV (Populismus) verwendet.

Der Replikationsansatz als Prüfverfahren

Die von uns verwendeten Datenfusionsverfahren sind genauer genommen Spezialfälle der Datenimputation. Das wird bei dem Verfahren mit den Zufallsstichproben genauer sichtbar als bei der Verknüpfung von zwei unabhängigen Stichproben. Hier wurden die Informationen der spezifischen Variablen aus der zweiten Datenhälfte herausgelöscht und somit als fehlende Werte definiert. Im nächsten Schritt werden aus den bestehenden Werten der ersten Datenhälfte auf Basis der festgelegten gemeinsamen Variablen die imputierten Werte berechnet. Die eigentliche Datenfusion wurde damit über die Imputation durch Einbindung des R-simulationPackages (Van der Loo 2022) durchgeführt.

Um die Qualität der Datenfusionen genauer zu testen können zwei Perspektiven eingenommen werden: Unsere Hauptperspektive folgt dem *Replikationsansatz*, da wir wissen wollen, wie hoch der Anteil an replizierten Effekten bei den fusionierten Datensätzen ist. Der *Standardansatz* folgt den üblichen Prüfverfahren bei Datenfusionen oder multiplen Imputationen (z.B. über Ermittlung des Root Mean Square Errors (RMSE) oder Mean Absolute Percentage Error (MAPE)), indem gemessen wird, wie stark die Neuberechneten Werte der betrachteten Variablen von den tatsächlichen Werten der Originalvariablen abweichen oder ob das Grundrauschen insgesamt erhalten bleibt (vgl. Karunasingha 2022; McKenzie 2011). Derartige Berechnungen werden für den

vorliegenden Zweck nicht durchgeführt, sollten aber im Kontext einer ganzheitlichen Betrachtung bei zukünftigen Untersuchungen dieser Art integriert werden.

Beim Replikationsansatz orientieren wir uns grob an den v.a. in der Psychologie üblichen Replikationskriterien. Auch wenn es noch keinen Standardansatz für die Evaluation von Replikationen gibt (Open Science Collaboration 2015), können wir uns an den Prüfgrößen bei der größten Replikationsstudie orientieren: Hier wurde die Reproduzierbarkeit anhand von Signifikanz- und P-Werten, Effektgrößen, subjektiven Einschätzungen von Replikationsteams und anhand von Meta-Analysen der Effektgrößen gemessen (Open Science Collaboration 2015). Eine einfache Methode zur Bewertung der Replikation sei beispielsweise die Prüfung, ob die Replikation eine statistisch signifikante Wirkung ($p < .05$) in der gleichen Richtung wie die Originalstudie zeigt (Open Science Collaboration 2015; vgl. Patil et al. 2016). Die Reproduzierbarkeit einer Studie ist die Wahrscheinlichkeit, dass eine neue Analyse, die auf dieselbe wissenschaftliche Frage abzielt, ein konsistentes Ergebnis liefert. (Leek & Peng: 2015, p. 1645; siehe auch Palti et al. 2016: 540; Asendorpf et al., 2013; Ioannidis, 2005). Dabei werden die Anteile an replizierten Effekten im Vergleich zwischen den Datenfusionsverfahren betrachtet. Neben dem reinen Signifikanzvergleich werden aber im Folgenden auch strengere Kriterien anhand von regressionspezifischen Kennwerten angesetzt. Bei den im Fokus stehenden repräsentativen Erhebungen mit großen Fallzahlen sind stabilere Ergebnisse zu erwarten als bei den in der Psychologie üblichen Experimenten oder bei den sog. korrelativen Studien mit jeweils niedrigeren und schwankenden Stichprobengrößen (vgl. Patil et al. 2016: 540). Daher werden neben dem 5%-Kriterium, noch das 1%-Kriterium und die exakte Replikation (basierend auf zwei Nachkommastellen) überprüft, auch um differenzierendere Ergebnisse zu bekommen. Auf die standardisierten Beta-Koeffizienten², die Signifikanz des Regressionsmodells und den Abgleich der Konfidenzintervalle mit dem Originalergebnis (Open Science Collaboration 2015) wird an dieser Stelle verzichtet. Es ist nicht nötig, um das Prinzip des Prüfverfahrens zu zeigen.

Ergebnisse der Replikation

Wie der Anteil an replizierten Effekten berechnet und zu einem Score verdichtet wird, ist in *Tab. 2* dargestellt. Die Ergebnisse der Variablen 1 bis 5 und 23 bis 26 sind dargestellt. Die Systematik folgt derjenigen in der Tabelle zur Haupthypothese, deren Ergebnisse in der ersten Variablenzeile wiederzufinden sind.

Für 26 unabhängige Variablen wurden jeweils bivariate Regressionen auf den Populismus-Index berechnet. Eingetragen sind die berechneten b-Koeffizienten, die aufgrund der einheitlichen „0 bis 1“-Skalierung aller unabhängigen Variablen auch miteinander verglichen werden können. Zuerst wurden alle Regressionen mit dem Allbus 2018 berechnet und mit den publizierten Ergebnissen abgeglichen. Da nur 25 der 26 publizierten Regressionen exakt repliziert werden konnten, dient im weiteren Verlauf die Replikation mit Allbus 2018 als Referenz. Die zentrale Prüfgröße ist die *proportion of replicated effects (pore-Score)*. Der pore-Score wird in drei Varianten eingeführt. Zuerst wird die Anzahl an exakt replizierten Effekten berechnet. Als exakt repliziert gilt ein b-Koeffizient, wenn er bis auf zwei Nachkommastellen genau repliziert wird. Dieses Kriterium wird

² Nach Kohler und Kreuter (2017: 288) ist die Verwendung von Beta-Koeffizienten bei den (hier überwiegend) verwendeten dichotomen Variablen nicht zulässig (für einen Überblick siehe z. B. Wolf und Best 2010: 625–628).

als zu streng angesehen. Das zeigt sich auch bei den beiden Allbuszufallsstichproben, die ohne Fusionen, aber mit einer kleineren Fallzahl auskommen müssen. Eine exakte Replikation gelingt nur bei ca. 42% (Hälfte 1) bzw. 35% (Hälfte 2). Bei den fusionierten Daten liegt der pore-Score bei KNN etwas niedriger mit 23%, bei PMM und SHD mit 42% und 31% in einem ähnlichen Bereich wie bei den Allbus-Hälften. Die Berechnung exakter Replikationen bietet den Vorteil einer deutlich differenzierenden Einordnung der Verfahren. Die beiden anderen Varianten folgen der Logik der Hypothesentests auf der Basis der 1%- bzw. 5%-Level. Wenn die berechneten Ergebnisse in einem Raum zwischen einem Prozentpunkt über oder unter dem Koeffizienten liegen, gilt das Ergebnis als repliziert auf dem 1%-Level. Auf dem 1%-Level können bei den beiden Hälften ca. 80% repliziert werden. Bei den fusionierten Daten sind es 70% bis 80%. Liegen die berechneten Ergebnisse in einem Raum zwischen fünf Prozentpunkten über oder unter dem Koeffizienten, gilt das Ergebnis als repliziert auf dem 5%-Level. Die Berechnungen der pore-Scores auf dem 5%-Level führten fast alle zu einer 100-prozentigen Replikation, lediglich bei PMM konnte einer von 26 Effekten (=96%) nicht repliziert werden.

Tabelle 2: Regression von unabhängigen Variablen (uV) auf den Populismus (aV): Berechnung des pore-Scores für b-Koeffizienten im Vergleich zwischen Originaldatensatz und drei fusionierten Datensätzen

| b-Koeffizienten | | | | | | | | | | | | | |
|---|------------------------|--|---|--------------------------------|--|-----------------------------|--|---------------|---|---------------|---|---------------|---|
| | Westie 2020: 226 | Replika- tion mit Allbus 2018 | Differenz Replikation mit Allbus 2018 zu Westie 2020 | Allbus 2018, Hälfte 1 | Diff. Hälfte 1 zu Replika- tion mit Allbus 2018 | Allbus 2018, Hälfte 2 | Diff. Hälfte 2 zu Replika- tion mit Allbus 2018 | KNN | Diff. KNN zu Replika- tion mit Allbus 2018 | PMM | Diff. PMM zu Replika- tion mit Allbus 2018 | SHD | Diff. SHD zu Replika- tion mit Allbus 2018 |
| gewichtet | | | | | | | | | | | | | |
| 1 Wissensindex 1 (19 Items) | -0,22 | -0,22 | 0,00 | -0,21 | 0,01 | -0,23 | -0,01 | -0,20 | 0,02 | -0,25 | -0,03 | -0,23 | -0,01 |
| 2 Ohne und Hauptschulabschluss | 0,13 | 0,13 | 0,00 | 0,13 | 0,00 | 0,12 | -0,01 | 0,13 | 0,00 | 0,13 | 0,00 | 0,14 | 0,01 |
| 3 Mittlere Reife, noch Schüler | 0,06 | 0,06 | 0,00 | 0,06 | 0,00 | 0,07 | 0,01 | 0,07 | 0,01 | 0,06 | 0,00 | 0,08 | 0,02 |
| 4 Fachhochschulreife | -0,06 | -0,06 | 0,00 | -0,07 | -0,01 | -0,05 | 0,01 | -0,07 | -0,01 | -0,06 | 0,00 | -0,09 | -0,03 |
| 5 Hochschulreife | -0,15 | -0,15 | 0,00 | -0,15 | 0,00 | -0,15 | 0,00 | -0,15 | 0,00 | -0,15 | 0,00 | -0,16 | -0,01 |
| 6 bis 22 [...] | | | | | | | | | | | | | |
| 23 (R) bis 4999 | 0,06 | 0,06 | 0,00 | 0,04 | -0,02 | 0,08 | 0,02 | 0,04 | -0,02 | 0,04 | -0,02 | 0,06 | 0,00 |
| 24 5.00 bis 49.999 | 0,00 | 0,00 | 0,00 | 0,02 | 0,02 | -0,01 | -0,01 | 0,02 | 0,02 | 0,02 | 0,02 | 0,01 | 0,01 |
| 25 50.000 bis 499.999 | -0,02 | -0,02 | 0,00 | -0,02 | 0,00 | -0,03 | -0,01 | -0,02 | 0,00 | -0,02 | 0,00 | -0,02 | 0,00 |
| 26 500.000 und größer | -0,05 | -0,05 | 0,00 | -0,06 | -0,01 | -0,05 | 0,00 | -0,07 | -0,02 | -0,06 | -0,01 | -0,06 | -0,01 |
| number of exactly replicated effects | | | 25 | 11 | | 9 | | 6 | | 11 | | 8 | |
| pore-Score: exactly replicated | | | 96,15 % | 42,31% | | 34,62% | | 23,08% | | 42,31% | | 30,77% | |
| number of replicated effects (1%-Level) | | | 26 | 20 | | 22 | | 18 | | 18 | | 21 | |
| pore-Score: 1%-Level | | | 100% | 76,92% | | 84,62% | | 69,23% | | 69,23% | | 80,77% | |
| number of replicated effects (5%-Level) | | | 26 | 26 | | 26 | | 26 | | 25 | | 26 | |
| pore-Score: 5%-Level | | | 100% | 100% | | 100% | | 100% | | 96,15% | | 100% | |

Anmerkungen: KNN = K-Nearest Neighbor; PMM = Predictive Mean Matching; SHD = Sequential Hot Deck Imputation; pore = proportion of replicated effects

Ähnliche Berechnungen und Vergleiche wurden auch mit den ermittelten Signifikanzniveaus (Anzahl Sterne) und den berechneten r^2 -Werten durchgeführt. *Tabelle 3* zeigt die ermittelten pore-Scores für die drei unterschiedlichen Maße b-Koeffizienten, r^2 -Werte und Signifikanzen im Vergleich.

Bei den r^2 -Werten wurden die drei pore-Scores in identischer Form berechnet wie bei den b-Koeffizienten, weil beide Maße zwischen 0 und 1 liegen. Bei den Signifikanzen wurden nur zwei Formen berechnet: Neben dem Anteil an exakten Replikationen wurde berechnet, ob Signifikanzen richtig erkannt wurden oder nicht. Dabei wurde nicht mehr nach dem genauen Signifikanzniveau geprüft.

Tabelle 3: Regression von unabhängigen Variablen (uV) auf den Populismus (aV): pore-Scores für b-Koeffizienten, r²-Werten und signifikanz im Vergleich zwischen Originaldatensatz und drei fusionierten Datensätzen

| | gewichtet | Differenz Replikation mit Allbus 2018 zu Westle 2020 | Diff. Hälfte 1 zu Replikation mit Allbus 2018 | Diff. Hälfte 2 zu Replikation mit Allbus 2018 | Diff. KNN zu Replikation mit Allbus 2018 | Diff. PMM zu Replikation mit Allbus 2018 | Diff. SHD zu Replikation mit Allbus 2018 |
|-----------------|--------------------------------|---|---|---|--|--|--|
| b-Koeffizienten | pore-Score; exactly replicated | 96,15 | 42,31 | 34,62 | 23,08 | 42,31 | 30,77 |
| | pore-Score; 1%-Level | 100,00 | 76,92 | 84,62 | 69,23 | 69,23 | 80,77 |
| | pore-Score; 5%-Level | 100,00 | 100,00 | 100,00 | 100,00 | 96,15 | 100,00 |
| r ² | pore-Score; exactly replicated | 96,15 | 76,92 | 69,23 | 69,23 | 34,62 | 61,54 |
| | pore-Score; 1%-Level | 100,00 | 100,00 | 100,00 | 92,31 | 69,23 | 96,15 |
| | pore-Score; 5%-Level | 100,00 | 100,00 | 100,00 | 100,00 | 88,46 | 100,00 |
| Signifikanz | pore-Score; exactly replicated | 96,15 | 76,92 | 76,92 | 80,77 | 84,62 | 100,00 |
| | pore-Score; significance | 100,00 | 84,62 | 92,31 | 88,46 | 96,15 | 100,00 |

Anmerkungen: KNN = K-Nearest Neighbor; PMM = Predictive Mean Matching; SHD = Sequential Hot Deck Imputation; pore = proportion of replicated effects

Zwischenfazit und Ausblick

Auf die theoriegeleitete Perspektive zurückkommend zeigt sich, dass die Hypothesentests mit den fusionierten Daten zu 90% bis 100% über alle Koeffizienten hinweg repliziert werden konnten. Dabei unterscheiden sich diese Ergebnisse kaum von Ergebnissen bei den beiden Allbus-Hälften, die sich von Allbus 2018 nur bezogen auf die halbierte Fallzahl unterscheiden. Beide Hälften können mit jeweils ca. 1700 Fällen pro Hälfte immer noch als fallzahlenstark bezeichnet werden. Das spricht einerseits für die Brauchbarkeit der Datenfusionsverfahren, andererseits für die Nutzbarkeit des pore-Scores. Mit Hilfe einer Vielzahl von theoriegeleiteten Verfahren zur Prüfung von Datenfusionen kann sich der pore-Score bewähren, wenn ausreichend Referenzdaten vorhanden sind.

Literatur

- ALLBUS 2018. GESIS Datenarchiv, Köln. ZA5270 Datenfile Version 2.0.0 (2019).
- Andridge, R.R., Little, R.J. (2010). A review of hot deck imputation for survey non-response. *International statistical review*, 78(1), 40-64.
- Asendorpf, J. et al. (2013). Recommendations for Increasing Replicability in Psychology. *European Journal of Personality*, 27, 108-119. 10.1002/per.1919.
- Axenfeld, J. B., Blom, A. G., Bruch, Ch. & Wolf, Ch. (2022). Split Questionnaire Designs for Online Surveys: The Impact of Module Construction on Imputation Quality. *Journal of Survey Statistics and Methodology*, 10 (5): 1236–262. <https://doi.org/10.1093/jssam/smab055>.
- Bacher, J., & Prandner, D. (2018). Datenfusion in der sozialwissenschaftlichen Wahlforschung – Begründeter Verzicht oder ungenutzte Chance? Theoretische Vorüberlegungen, Verfahrenüberblick und ein erster Erfahrungsbericht. In: *Austrian Journal of Political Science*, 47(2), pp. 61-76.
- Diekmann, A., Hadjar, A., Kurz, K., Rosar, U., Wagner, U. & Westle, B. (2019). Allgemeine Bevölkerungsumfrage der Sozialwissenschaften. GESIS - Leibniz-Institut für Sozialwissenschaften (2019). Allgemeine Bevölkerungsumfrage der Sozialwissenschaften ALLBUS 2018. GESIS Datenarchiv, Köln. ZA5270 Datenfile Version 2.0.0, <https://doi.org/10.4232/1.13250>
- Gower, J. C. (1971). A general coefficient of similarity and some of its properties. In: *Biometrics*, pp. 857-871.
- Hagenah, J. & Meulemann, H. (2009). The Analytical Potentials of Trend Survey Data from Market Research. The Case of German Media Analysis Data. *Historical Social Research (HSR)/ Historische Sozialforschung* 34/1, 49-61.
- Ioannidis, J.P.A. (2005). Why Most Published Research Findings Are False. *PLoS Med* 2(8): e124. <https://doi.org/10.1371/journal.pmed.0020124>
- Jadhav, A., Pramod, D., . & Ramanathan, K. (2019). Comparison of Performance of Data Imputation Methods for Numeric Dataset. *Applied Artificial Intelligence*, 33(10), 913–933.
- Karunasingha, D.S.K. (2022). Root mean square error or mean absolute error? Use their ratio as well. *Information Sciences*, 585, 609-629.
- Kerr, N.L. (1998). HARKing: hypothesizing after the results are known. *Personality and Social Psychology Rev.* 1998; 2(3),196-217. doi: 10.1207/s15327957pspr0203
- Khaleghi, B., Khamis, A., Karray, F.O. & Razavi, S.N. (2013). Multisensor Data Fusion: A Review of the State-of-the-Art. *Information Fusion*, 14, 28-44. <https://doi.org/10.1016/j.inffus.2011.08.001>
- Kohler, U. & Kreuter, F. (2017). *Datenanalyse mit Stata. Allgemeine Konzepte der Datenanalyse und ihre praktische Anwendung*. Berlin, DeGruyter.
- Leek, J., & Peng, R. (2015). Statistics: P values are just the tip of the iceberg. *Nature* 520, 612 (2015). <https://doi.org/10.1038/520612a>
- Little, R. J. A. (1988). Missing data adjustments in large surveys. In: *Journal of Business & Economic Statistics*, 6(3), pp. 287-296.
- McKenzie, J. (2011). Mean absolute percentage error and bias in economic forecasting. *Economics Letters*, Volume 113(, Issue 3), 259-262.
- Moosdorf, D., Schnaudt, C., Tausendpfund, M., & Westle, B. (2020). Messung politischen Wissens. In: M. Tausendpfund & M.,B. Westle, B. (eds), *Politisches Wissen in Deutschland*.

- Politisches Wissen. Springer VS, Wiesbaden: Springer VS. https://doi.org/10.1007/978-3-658-30492-8_3
- Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science* 349, aac4716 (2015). DOI: 10.1126/science.aac4716
- Patil, P., Peng, R. D., & Leek, J. T. (2016). What Should Researchers Expect When They Replicate Studies? A Statistical View of Replicability in Psychological Science. *Perspectives on Psychological Science*, 11(4), 539–544. <https://doi.org/10.1177/1745691616646366>
- Rhaudatunnisa, T., & Wilantika, N. (2022). Performance Comparison of Hot-Deck Imputation, K-Nearest Neighbor Imputation, and Predictive Mean Matching in Missing Value Handling, Case Study: March 2019 SUSENAS Kor Dataset. In: *Proceedings of The International Conference on Data Science and Official Statistics, 2021(1)*, 753–770.
- Van der Loo, M. (2022). Package *simputation*. R package version 0.2.8, retrieved from <https://cran.r-project.org/web/packages/simputation/index.html>.
- Westle, B. (2020). Schützt politisches Wissen vor Populismus?. In: Tausendpfund, M., Westle, B. (eds) *Politisches Wissen in Deutschland. Politisches Wissen*. Springer VS, Wiesbaden. https://doi.org/10.1007/978-3-658-30492-8_7
- Wolf, C. & Best, H. (2010) (Hrsg.). *Handbuch der sozialwissenschaftlichen Datenanalyse*. Springer VS, Wiesbaden.
- Zhang, S. (2012). Nearest neighbor selection for iteratively kNN imputation. In: *Journal of Systems and Software*, 85(11), pp. 2541-2552.
- zpid (2023). PreReg: Preregistration in Psychology. Trier: zpid – Leibniz Institut für Psychologie. <https://prereg-psych.org/index.php/rpp>.

Datenfusionen zielen darauf ab, durch die Kombination mehrerer Datenquellen vollständige Datensätze zu gewinnen. Im Mittelpunkt des Beitrags steht die Einführung eines Prüfverfahrens zur Qualitätsmessung von Datenfusionen. Dafür erfolgt zunächst eine kurze methodische Vorstellung der verwendeten Datenfusionsverfahren. Anschließend wird am Beispiel der Studie „Schützt politisches Wissen vor Populismus?“ von Westle (2020) eine Replikation der Hypothesenprüfung mit verschiedenen fusionierten Datensätzen vorgenommen. Die Ergebnisse können als Beleg für die Brauchbarkeit fusionierter Datensätze für komplexere Kausalanalysen gelten und zeigen die Tauglichkeit verschiedener Verfahren der Datenfusion im Vergleich.