Weiand, Severin V.

Article

Vergleich von Record-Linkage-Methoden anhand der Mikrosimulation eines bundesweiten Bildungsverlaufsregisters

WISTA - Wirtschaft und Statistik

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

Suggested Citation: Weiand, Severin V. (2025) : Vergleich von Record-Linkage-Methoden anhand der Mikrosimulation eines bundesweiten Bildungsverlaufsregisters, WISTA - Wirtschaft und Statistik, ISSN 1619-2907, Statistisches Bundesamt (Destatis), Wiesbaden, Vol. 77, Iss. 4, pp. 95-105

VERGLEICH VON RECORD-LINKAGE-METHODEN ANHAND DER MIKRO-SIMULATION EINES BUNDESWEITEN BILDUNGSVERLAUFSREGISTERS

Severin V. Weiand

Schlüsselwörter: Registermodernisierung – Zensus – Bildungsmodul – Linkage-Bias – Datenqualität

ZUSAMMENFASSUNG

Damit ein Register wie das geplante bundesweite Bildungsverlaufsregister optimal genutzt werden kann, sind vorab die technischen Probleme zu klären, die bei der Zusammenführung der Daten einer Person aus verschiedenen Datenquellen, dem sogenannten Record-Linkage, auftreten können. Einige der dabei nötigen Entscheidungen lassen sich nicht anhand vorhandener empirischer Daten treffen. Eine Mikrosimulation kann in der Diskussion um ein bundesweites Bildungsverlaufsregister genutzt werden, um entsprechende Fragen zu beantworten. Die Auswirkungen verschiedener Record-Linkage-Methoden auf die Qualität des Registers werden hierzu bei einem simulierten Register getestet. Überdies wird untersucht, inwieweit ein Linkage-Bias durch die verschiedenen Methoden entsteht.

Keywords: *register modernisation – census – education module – linkage bias – data quality*

ABSTRACT

For optimal use of a register such as the planned national educational trajectory register, it is first necessary to address the technical problems that may occur when linking data that refer to the same person across different data sources, known as record linkage. Some of the necessary decisions cannot be taken based on existing empirical data. In the context of the discussion about a national educational trajectory register, a microsimulation can be used to clarify such issues. To this end, the effects of different record linkage methods on the quality of the register are tested using a simulated register. In addition, the extent to which linkage bias is created by the various methods is analysed.

Severin V. Weiand

hat Behavioural Data Science (M. A.) mit dem Schwerpunkt Survey Methodology und zuvor Soziologie (B. A.) sowie Informatik (B. Sc.) studiert. Seit 2024 ist er wissenschaftlicher Mitarbeiter und Doktorand am Lehrstuhl für Methoden der empirischen Sozialforschung an der Universität Duisburg-Essen.

Für seine 2022 angefertigte Bachelorarbeit „Vergleich von Record-Linkage Methoden anhand der Mikro-Simulation eines bundesweiten Schülerregisters“, die dieser Artikel vorstellt, erhielt er den Wissenschaftlichen Nachwuchspreis „Statistical Science for the Society“ 2024. Die Arbeit wurde betreut von Prof. Dr. Rainer Schnell an der Universität Duisburg-Essen.

1

Einleitung

Im Zuge der Registermodernisierung und des geplanten Registerzensus gewinnen Verfahren zur Verknüpfung von Registern zunehmend an Bedeutung (Haußmann, 2018). Neben der Führung einer Identifikationsnummer (ID) als Verwaltungsmerkmal stehen dabei auch Record-Linkage-Methoden im Fokus (Schnell, 2019). Hierbei werden Individualdaten anhand von Quasi-Identifikatoren, wie Name oder Geburtsdatum, verknüpft und damit zugehörige Datensatzeinträge identifiziert.

Ein zentrales Problem von Record-Linkage besteht darin, dass Quasi-Identifikatoren nicht eindeutig sind und Fehler enthalten können (Herzog und andere, 2007). Entsprechende Methoden müssen dies berücksichtigen, um eine qualitativ hochwertige und verzerrungsfreie Verknüpfung zu gewährleisten. Andernfalls können die verknüpften Datensätze nicht das Qualitätskriterium der amtlichen Statistik der Kohärenz und Vergleichbarkeit erfüllen (Statistische Ämter des Bundes und der Länder, 2021).

Das Beurteilen der Verknüpfungsqualität von Realdaten ist jedoch schwierig, da hierfür die Korrektheit eines potenziellen Recordpaares bekannt sein muss (Herzog und andere, 2007). Um ein Record-Linkage-Verfahren zu implementieren, insbesondere für große Projekte, wie es bei der Registermodernisierung der Fall ist, sollte daher das verwendete Verfahren zunächst getestet werden. Dies ist besonders bedeutsam, wenn vor der Einführung des Registers die benötigten Quasi-Identifikatoren festgelegt werden müssen und deren Menge aufgrund der Anforderung der Datensparsamkeit möglichst klein gehalten werden soll (Schnell, 2019).

Eine Möglichkeit, Testdaten zu gewinnen, ist die Mikrosimulation. Dabei handelt es sich um die Simulationen auf der Ebene einzelner Einheiten (Merkmalsträger), wie in diesem Fall Personen, über die Zeit (Li/O'Donoghue, 2013). Dies erlaubt auch den Test von Längsschnittverknüpfungen.

Konkret wurde in dieser Arbeit die Mikrosimulation eines bundesweiten Bildungsverlaufsregisters durchgeführt. Zunächst wird hierfür in Kapitel 2 der Hintergrund dieser Simulation beschrieben. Daran schließt sich eine kurze Darstellung wesentlicher Elemente des Simulationsprozesses an. Die erzeugten simulierten Daten wurden danach mithilfe einer Vielzahl von Record-Linkage-Verfahren verknüpft, wie Kapitel 4 erläutert. Abschließend erfolgt eine Analyse der Verknüpfungsergebnisse hinsichtlich der Linkage-Qualität, des Linkage-Bias und der möglichen Auswirkungen auf Kohärenz und Vergleichbarkeit der Daten.

2

Hintergrund

Im Zuge der Umstellung des Zensus auf ein registerbasiertes Verfahren diskutieren Bund und Länder seit längerer Zeit über die Einführung eines Bildungsverlaufsregisters (Giar und andere, 2023). Ziel eines solchen Registers ist, Bildungsverläufe auf Individualebene in Form eines Längsschnittregisters abzubilden. Im Zuge des geplanten Registerzensus wird dieses Register als Bildungsmodul benötigt, um die Lieferverpflichtungen gegenüber der Europäischen Union für den Zensus zu erfüllen (Gawronski, 2020). Überdies wird erwartet, dass das Register bundesweite Analysen des Schulsystems und Bildungserfolgs erlaubt, die aufgrund sehr hoher Kosten durch Erhebungen kaum zu erreichen sind. Insbesondere die Fähigkeit, den Bildungserfolg bestimmter Subpopulationen, wie Migranten, zu analysieren, ist dabei eine konkrete Erwartung an das Register. Auch wird gefordert, dass Bildungsübergänge über Bundesländer hinweg erforscht werden können (RatSWD, 2022).

Zentral für das Erfüllen dieser Erwartungen an ein Bildungsverlaufsregister ist ein funktionierendes Identitätsmanagement. Bisherige Überlegungen sehen vor, dass das Bildungsverlaufsregister durch die Verknüpfung anhand von Record-Linkage-Verfahren erstellt werden soll (Giar und andere, 2023). Von besonderer Bedeutung ist dabei, dass die Verknüpfung zu keinem Linkage-Bias führt. Bestimmte Subpopulationen sollen daher keinen signifikant anderen Verknüpfungserfolg haben als der Rest der Population. Andernfalls bestehen Unterschiede in der Qualität der abgebildeten Bildungsverläufe (Schnell, 2022).

Um die Möglichkeiten des Verwendens von Record-Linkage-Verfahren zur Erstellung des Bildungsverlaufsregisters zu prüfen, beauftragte das Bundesministerium für Bildung und Forschung Prof. Dr. Rainer Schnell damit,

eine Simulation des Registers durchzuführen. Insbesondere sollten die notwendigen Quasi-Identifikatoren für ein bundesweites Bildungsverlaufsregister untersucht werden. Der Bericht (Schnell, 2022) enthält die Details der Simulation, eine knappe Zusammenfassung findet sich bei Schnell und Weiand (2023). Die auf der Simulation aufbauende Bachelorarbeit (Weiand, 2022) sollte anhand der simulierten Datensätze einen detaillierten Vergleich verschiedener Record-Linkage-Verfahren durchführen. Entsprechend wird die Mikrosimulation im Folgenden nur in ihren Grundzügen erläutert.

3 Beschreibung der Mikrosimulation

Das Register wurde anhand von jährlichen Querschnitten für eine Periode von zehn Jahren simuliert. Die einzelnen Querschnitte wurden dabei mit dem jeweiligen Vorjahresergebnis verknüpft. Wie viele Mikrosimulationen bestand auch diese Simulation somit aus zwei Phasen (Li/O'Donoghue, 2013). Zunächst wurde ein Startdatensatz erstellt, der den Zustand des Registers zur virtuellen Einführung im Jahr 2000 möglichst genau darstellte und größtenteils durch einen Vorlauf der im Folgenden dargestellten Simulationsroutine gebildet wurde. Anschließend erfolgte die eigentliche Mikrosimulation, bei der – basierend auf dem Vorjahresstand – die jährlichen Veränderungen der Quasi-Identifikatoren simuliert wurden. Jährlich wird das Register zudem um weitere Bildungsteilnehmende erweitert. Da keine Personen gelöscht werden, ist das Register somit kumulativ. Das Register umfasst zum Startjahr etwa 18,9 Millionen und zum Endjahr (2009) etwa 27 Millionen simulierte Personen.

Für jede Person wurden die Quasi-Identifikatoren Vorname, Nachname, Geburtsdatum, Geschlecht und Geburtsort simuliert. Quasi-Identifikatoren folgen keinen einfachen Verteilungen und korrelieren zumeist miteinander. Beispielsweise ist die Namensgebung regional und zeitlich abhängig. Dies führt dazu, dass Kombinationen von Quasi-Identifikatoren häufiger auftreten, als es unter Unabhängigkeit zu erwarten ist. Dies wiederum erschwert die Verknüpfung, da diese Personen schwerer zu unterscheiden sind. Kommt es darüber hinaus zu Fehlern bei der Eingabe, erschwert dies die Verknüpfung zusätzlich. Entsprechend werden diese Zusammenhänge möglichst detailliert simuliert, um möglichst realistische Verknüpfungsergebnisse zu erhalten.

Es gilt zu beachten, dass der Detailgrad einer Mikrosimulation immer von den vorhandenen Daten und Ressourcen abhängt. Entsprechend müssen einige reale Prozesse vereinfacht werden. In diesem Fall waren insbesondere die Ausreißer im Bildungssystem schwer zu modellieren. Der Bildungsverlauf wurde daher relativ linear modelliert, wie sich im Folgenden zeigen wird.

Zur Vereinfachung der simulierten Prozesse wird insbesondere zwischen vier Personengruppen unterschieden:

1. Schülerinnen und Schüler,
2. Studierende,
3. Auszubildende und
4. Erwachsene.

Innerhalb dieser Gruppen werden zumeist unterschiedliche, feste Wahrscheinlichkeitsverteilungen angenommen. Ferner wird zwischen in Deutschland geborenen Personen (Einheimische) sowie Migrantinnen und Migranten unterschieden.

3.1 Simulation der Personenmerkmale

Anhand der Geburtsquote wurde die Anzahl der Kinder ermittelt, die in einem bestimmten Jahr eingeschult wurden. Für Einheimische wurden Monats- und Wochentagsverteilungen der Geburtsdaten berücksichtigt. Zudem wurde auch der Anteil der Mehrlinge (Zwillinge und Drillinge) berücksichtigt, die bei der Verknüpfung besondere Probleme hervorrufen können, da sie sich zumeist nur in einem der erfassten Quasi-Identifikatoren unterscheiden. Ferner wurde angenommen, dass jährlich 49 700 im Ausland geborene Schülerinnen und Schüler in das Schulsystem eintreten. Proportional zur Anzahl gemeldeter ausländischer Studierender an den deutschen Hochschulen wurde zudem die Anzahl der Neuzugänge an den Hochschulen bestimmt.

Der Geburtsort wurde proportional zur Größe der Gemeinden zugeordnet. Der Anteil der Migrantinnen und Migranten wurde über den Ausländeranteil innerhalb einer Gemeinde bestimmt sowie über die Anzahl

ausländischer Studierender. Da für Migrantinnen und Migranten keine detaillierten Daten über Geburtsorte öffentlich verfügbar sind, wurde den simulierten Migrantinnen und Migranten zufällig ein Geburtsort im Ausland zugeordnet.

Häufige Namenskombinationen, beispielsweise Thomas Müller, sind schwerer zu verknüpfen; daher galt es, die Homogenität der aktuell vergebenen Namen bei Kindern möglichst detailliert nachzubilden. Hierzu wurde in Kooperation mit dem Institut für Bildungsmonitoring und Qualitätsentwicklung Hamburg eine verschlüsselte Namensverteilung nach Geschlecht der Hamburger Schülerschaft aus dem Zentralen Schülerregister verwendet. Diese Verteilung wurde auf Basis der Häufigkeiten einzelner Namensteile auf ganz Deutschland skaliert. Hierdurch wurden etwa 10 Millionen verschiedene Namen für die etwa 27 Millionen Personen im Register generiert.

Die Simulation beginnt mit der Einschulung. Nach Beendigung der Schulzeit beginnen Schülerinnen und Schüler eine Ausbildung oder ein Studium. Nach diesem Bildungsabschnitt endet die Bildungslaufbahn der entsprechenden Personen. Zudem können Personen nach Beendigung ihrer Schulzeit ihre Bildungslaufbahn beenden. Unterschiedliche Schul-, Ausbildungs- und Studienzeiten werden dabei berücksichtigt. So haben beispielsweise alle Schülerinnen und Schüler, die nach zehn Jahren die Schule verlassen (Haupt- oder Realschulabschluss) eine höhere Wahrscheinlichkeit, ihre Bildungslaufbahn zu beenden als Abiturientinnen und Abiturienten. Gleichzeitig können Personen, die nach zehn Jahren die Schule verlassen haben, kein Studium beginnen.

An allen Punkten der Bildungslaufbahn besteht zudem die Möglichkeit, dass eine Person durch Auswanderung oder Tod frühzeitig ausscheidet. Ferner wurden auch Ehen simuliert und die damit einhergehende mögliche Namensänderung.

3.2 Simulation von Dateneingabefehlern

Ein zentrales Problem beim Record-Linkage sind ungenaue oder veränderte Werte, die eine direkte Verknüpfung verhindern. Entsprechend ist die Fehlerroutine ein zentrales Element dieser Simulation. Dabei wird angenommen, dass Fehler durch eine Neuerfassung der Personendaten entstehen können. Diese Neuerfassung wird bei den Ereignissen ausgelöst, die im Bildungsverlauf eintreten können. Dies ist zunächst der Wechsel der Bildungseinrichtung nach Beendigung der Schulzeit sowie der Wechsel von der Grundschule zu einer weiterführenden Schule, der für alle Schülerinnen und Schüler nach vier Jahren angenommen wurde. Ferner können Neuerfassungen jederzeit durch Umzug und Heirat erfolgen.[1]

Da keine empirischen Daten über die mögliche Datenqualität eines Bildungsregisters zum Zeitpunkt der Simulation vorlagen, wurde die Datenqualität in Form von mehreren möglichen Fehlerquoten simuliert (0,1 %, 0,3 %, 0,7 % und 1 %). Dies erlaubt es, den Einfluss der Datenqualität auf die Verknüpfung abzuschätzen. Die Fehlerquote ist für alle Quasi-Identifikatoren gleich groß und legt den Anteil der Felder fest, die mit einem Fehler behaftet werden. Die Auswahl erfolgt dabei unabhängig. Bei einer Person können somit Fehler in mehreren Quasi-Identifikatoren auftreten.

Vor der Fehlergenerierung werden die Quasi-Identifikatoren der betroffenen Records auf ihren wahren Wert gesetzt. Vorhandene Fehler können somit auch korrigiert werden.

Für Textfelder (Vorname, Nachname und Geburtsort) wurden Tippfehler nach der Damerau-Levenshtein-Metrik simuliert (Damerau, 1964). Basierend auf Peterson (1986) wurden in 7,1 % der Fälle zwei Zeichen und in den restlichen Fällen ein Zeichen gelöscht, ersetzt, eingefügt oder vertauscht. Für alle weiteren Merkmale (Geburtstag, -monat, -jahr, Geschlecht) wurde das Feld mit einem anderen möglichen Wert ersetzt.

4 Record-Linkage-Verfahren

Zur Verknüpfung von Datensätzen anhand von nicht eindeutigen Quasi-Identifikatoren wurde eine Vielzahl von Methoden entwickelt (Christen, 2012). Die simulierten und korrumpierten Datensätze erlauben nun den Vergleich einiger dieser Methoden hinsichtlich der Linkage-Qualität und des entstehenden Linkage-Bias. In Form

1 Umzüge sind voneinander unabhängige Ereignisse, die abhängig von der Personengruppe simuliert wurden.

von zwei Linkage-Szenarien soll zudem die Auswirkung des Geburtsorts auf Qualität und Bias beurteilt werden.

Ein Record-Linkage-Verfahren klassifiziert ein Recordpaar entweder als Match oder Non-Match. Berücksichtigt man den wahren Zustand dieses Recordpaares, dann ergeben sich vier Kombinationen: True Positive (TP), True Negative (TN), False Positive (FP) und False Negative (FN). Hieraus können die beim Record-Linkage verwendeten Qualitätsmaße

$$Precision = \frac{TP}{TP + FP}$$

und

$$Recall = \frac{TP}{TP + FN}$$

berechnet werden (Christen, 2012). Üblicherweise wird in der amtlichen Statistik in Deutschland eine hohe Precision gefordert. Diese Anforderung wurde für die implementierten Verfahren übernommen.

Ein Linkage-Bias liegt vor, falls eine bestimmte Teilpopulation der Bevölkerung schlechter verknüpft werden kann als der Rest der Bevölkerung (Kvalsvig und andere, 2019). Dadurch weicht der wahre Populationswert für diese Subpopulation von dem durch die Verknüpfung ermittelten Populationswert ab. Zur Betrachtung des Bildungsverlaufsregisters wird der Mittelwert $\bar{x}$ der in allen Jahren richtig verknüpften Population mit dem wahren Mittelwert μ verglichen. Zur Quantifizierung dieser Abweichung wird der Standardized Bias verwendet (Collins und andere, 2001). Dieser ist definiert als

$$Bias = 100 \cdot \frac{\bar{x} - \mu}{s},$$

wobei s die Standardabweichung des durch die Verknüpfung ermittelten Populationswertes ist. Der Wert gibt somit die prozentuale Abweichung des Mittelwerts in Standardabweichungen an. Ein kritischer Effekt ist bei einer Abweichung von ±40 % zu erwarten (Collins und andere, 2001).

Record-Linkage-Methoden lassen sich zwischen deterministischen und probabilistischen Verfahren unterscheiden. Bei deterministischen Verfahren erfolgt ein exakter Abgleich einer Teilmenge von Merkmalen (Matchkeys). Bei probabilistischen Verfahren wird eine Match-Wahrscheinlichkeit durch den Vergleich aller Merkmale berechnet und mittels eines Schwellenwerts die Klassifizierung durchgeführt. Deterministische Verfahren liefern zumeist einen niedrigeren Recall als probabilistische Verfahren, benötigen dafür jedoch deutlich weniger Rechenleistung. Um den Rechenaufwand der probabilistischen Verfahren zu begrenzen, werden zumeist Blockingregeln eingeführt, die die Vergleichsmenge beschränken (Christen, 2012).

Die Verknüpfung des Bildungsverlaufsregisters erfolgte zweistufig. Im ersten Schritt fand eine deterministische Verknüpfung anhand aller verfügbaren Quasi-Identifikatoren statt (Exakter Matchkey). Hierdurch werden eindeutige Fälle aussortiert und der Rechenaufwand für das nächste Verfahren reduziert. Folgende Verfahren wurden im zweiten Schritt angewendet:

- Statistical Linkage Key (SLKL-581): Ein deterministischer Matchkey, gebildet aus dem zweiten und dritten Buchstaben des Vornamens, dem zweiten, dritten und fünften Buchstaben des Nachnamens, dem vollständigen Geburtsdatum und Geschlecht (Karmel, 2005).
- Probabilistisches Record-Linkage (PRL): Mithilfe eines Expectation-Conditional-Maximization-Algorithmus werden nach dem Fellegi-Sunter-Modell (Fellegi/Sunter, 1969) Match-Wahrscheinlichkeiten berechnet. Als Schwellenwert wurde 0,8 gewählt. Als Vergleichsfunktion für Textfelder wurde die Jaro-Winkler-Distanz verwendet (Herzog und andere, 2007).
- Cryptographic Long-term Keys (CLK): Alle Quasi-Identifikatoren werden in einen Bit-Vektor (Bloom-Filter) durch Aufteilen in N-Gramme codiert (Schnell und andere, 2011; Schnell und andere, 2009). Textfelder wurden als Bigramme und Zahlenfelder als Unigramme codiert. Es wurden Bloom-Filter mit 1 000 Bits erstellt und alle Merkmale mit jeweils zehn Hash-Funktionen verschlüsselt. Die Bloom-Filter wurden mithilfe von Multibit Trees indiziert (Kristensen und andere, 2010). Dabei wurde eine Tanimoto-Ähnlichkeit von 0,8 als Schwellenwert verwendet. Es wurde immer nur das erste relevante Token des Geburtsorts codiert.
- Multiple Matchkeys (MMK): Anhand der Parameter des Fellegi-Sunter-Modells werden redundanzfreie deterministische Matchkeys bestimmt (Randall und andere, 2019). Als Entscheidungskriterium für nicht eindeutige Matche wurde die Anzahl übereinstimmender Matchkeys genutzt (Schwellenwerte: mit Geburtsdatum = 1; ohne Geburtsdatum = 3).

Sofern der Geburtsort vorhanden war, wurden beim probabilistischen Record-Linkage folgende Blockingregeln sequenziell angewendet:

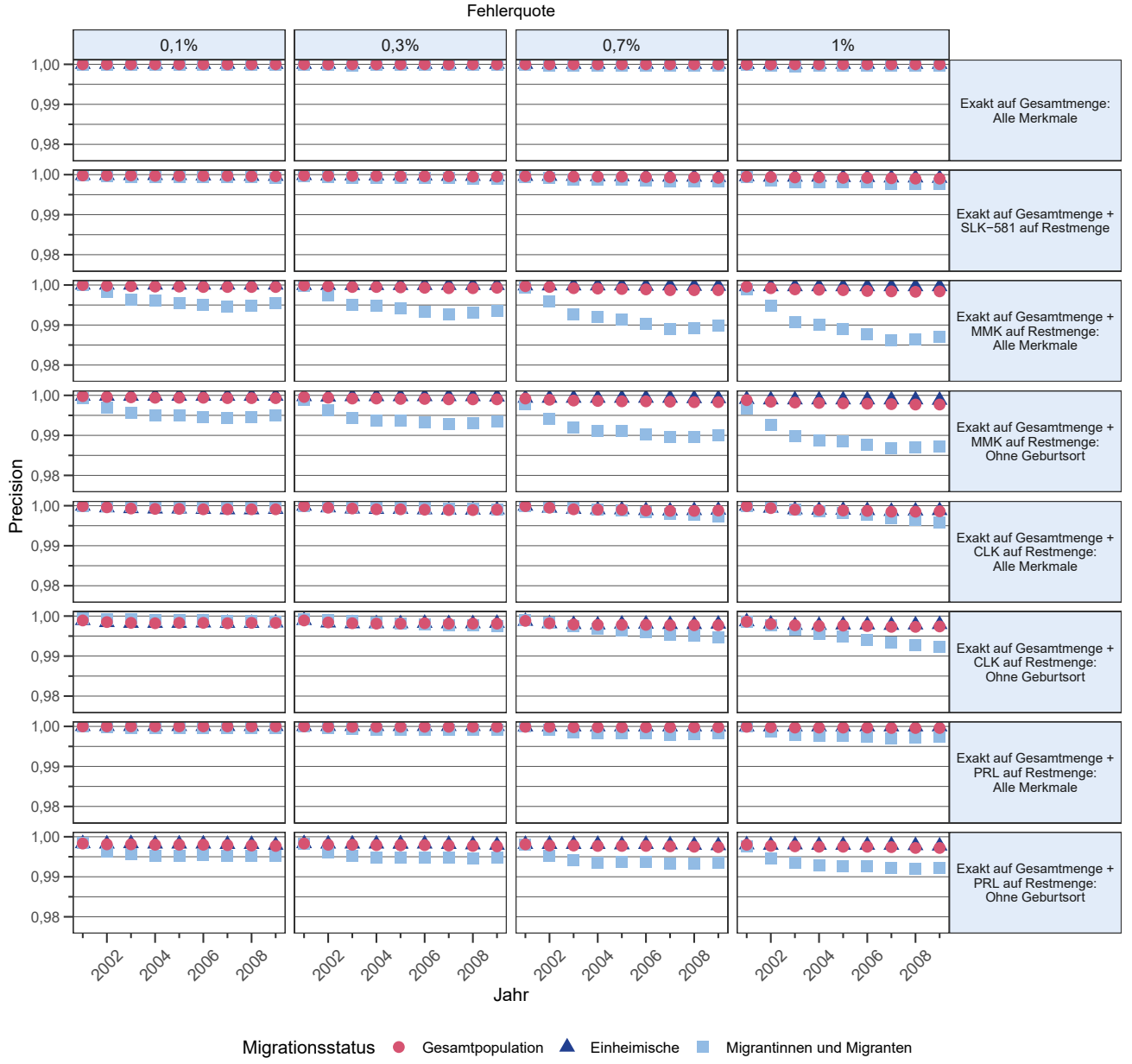
› mit Geburtsort: (1) Tag & Monat & Jahr, (2) Soundex (Vorname) & Soundex (Nachname), (3) Soundex (Vorname), (4) Geburtsort;

› ohne Geburtsort: (1) Tag & Monat & Jahr, (2) Soundex (Vorname) & Soundex (Nachname), (3) Soundex (Vorname) & Jahr.

Bei jeder angewendeten Blockingregel werden nur diejenigen Records einbezogen, die zuvor noch nicht verknüpft wurden.

Grafik 1

Precision der Verknüpfung aller einzelnen simulierten Jahre

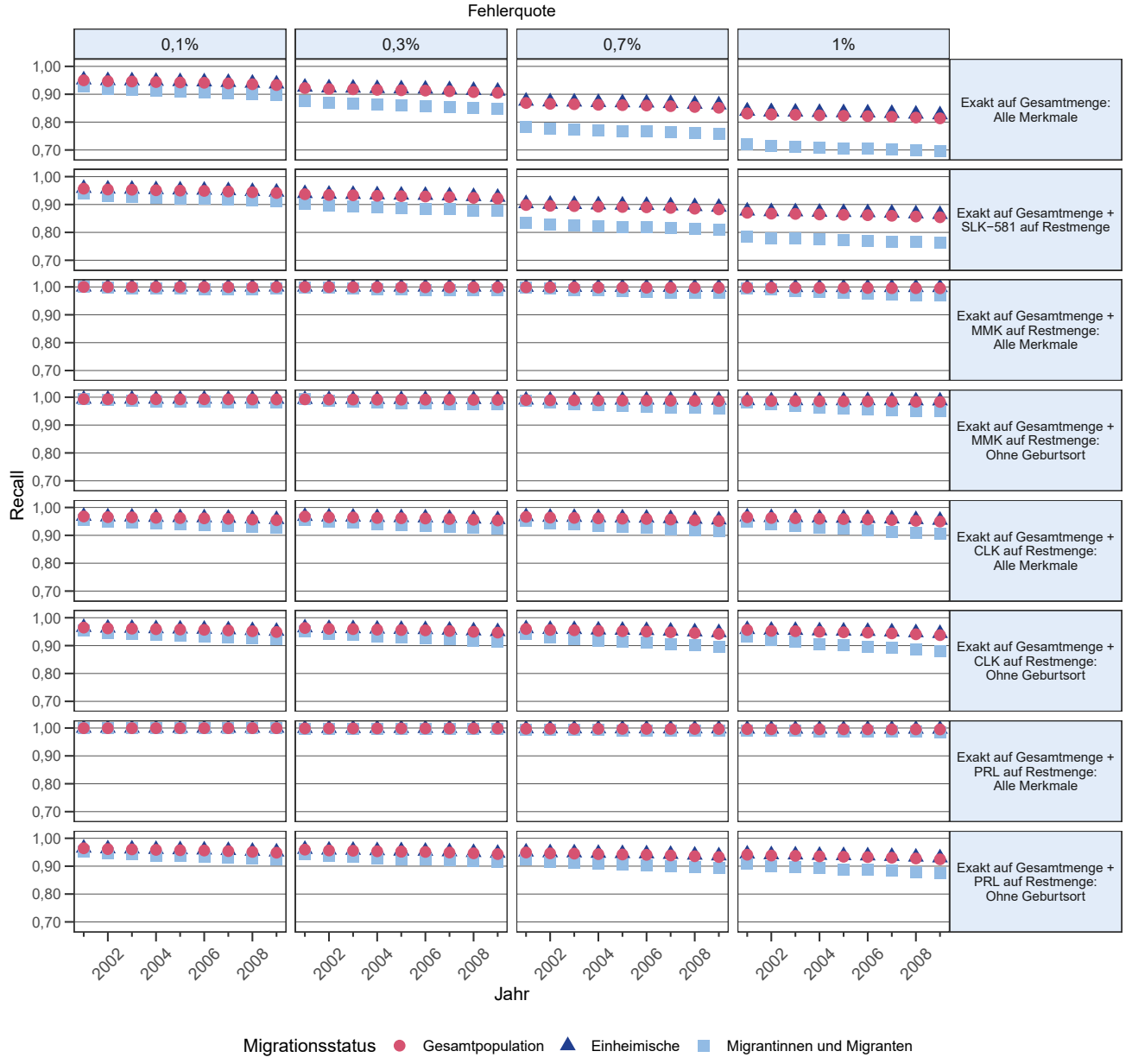


5

Ergebnisse

↘ Grafik 1 stellt die Precision der einzelnen Verfahren für alle Jahre dar, ↘ Grafik 2 den Recall. Hierbei wird zwischen den Gesamtergebnissen und den Ergebnissen nach Herkunft unterschieden. Wie erwartet, liefern die exakte Verknüpfung und deterministische Matchkeys eine sehr hohe Precision und einen niedrigen Recall. Mit einem Recall von über 0,8 bei allen Verknüpfungen zeigt sich, dass ein erheblicher Teil der Records durch die exakte Verknüpfung gefiltert werden konnte. Alle Verfahren weisen gemäß den Anforderungen eine hohe Precision auf. Insbesondere bei fehlendem Geburtsort sowie den MMK-Verfahren zeigt sich jedoch, dass die Precision bei der Verknüpfung von Migrantinnen und Migranten niedriger ist als bei Einheimischen.

Grafik 2
Recall der Verknüpfung aller einzelnen simulierten Jahre



Besonders bei der Betrachtung des Recalls zeigen sich unterschiedliche Verknüpfungserfolge bei Einheimischen sowie Migrantinnen und Migranten deutlich. Ausnahmen hiervon sind probabilistisches Record-Linkage (mit Verwendung des Geburtsorts) und die MMK-Verfahren. Diese Verfahren lieferten zudem die besten Verknüpfungsergebnisse, gefolgt vom probabilistischen Verfahren.

Ferner zeigt sich, dass die meisten Ergebnisse einen negativen Trend über die Zeit aufweisen. Ursache hierfür ist, dass sich der Datensatz vergrößert und – damit einhergehend – dass die Verwechslungswahrscheinlichkeit ansteigt. Überdies zeigt sich, dass die Linkage-Qualität mit zunehmender Fehlerquote sinkt. Das probabilistische Verfahren mit Geburtsort zeigt dabei den geringsten Einfluss der Fehlerquote auf die Linkage-Qualität. Ein sehr hoher Einfluss liegt bei den deterministischen Verfahren vor.

Der Standardized Bias einiger ausgewählter Werte ist in ↘ Tabelle 1 dargestellt. Auch hier weisen das probabilistische Verfahren mit Verwendung des Geburtsorts und die MMK-Verfahren den niedrigsten Bias auf, insbesondere wenn der Geburtsort vorhanden ist. Kritische Werte zeigen sich deutlich bei der Aufschlüsselung der Anzahl der Ehen nach Herkunft und Geschlecht. Insbesondere für weibliche Migranten zeigt sich ein kritischer Linkage-Bias bei den CLK-Verfahren und dem probabilistischen Record-Linkage ohne Geburtsort.

Tabelle 1
Standardized Bias ausgewählter Werte für die in allen Jahren korrekt verknüpften Records

Verfahren	Fehlerquote in %	Bildungsjahre insgesamt	Umzüge insgesamt	Ehen insgesamt	Ehen Einheimische, männlich	Ehen Migranten, männlich	Ehen Einheimische, weiblich	Ehen Migranten, weiblich
Exakt auf Gesamtmenge + SLK-581 auf Restmenge	0,1	– 11,501	– 3,882	– 25,401	– 3,497	– 5,774	**– 40,46**	**– 103,937**
	0,3	– 13,129	– 5,085	– 26,356	– 4,443	– 8,183	**– 41,538**	**– 107,727**
	0,7	– 16,528	– 7,847	– 28,448	– 6,455	– 13,649	**– 43,866**	**– 117,978**
	1	– 19,189	– 9,954	– 30,182	– 8,056	– 17,955	**– 45,861**	**– 127,077**
Exakt auf Gesamtmenge + MMK auf Restmenge: Alle Merkmale	0,1	0,198	0,199	– 0,286	– 0,014	– 0,531	– 0,304	– 1,518
	0,3	0,196	0,22	– 0,501	– 0,049	– 0,797	– 0,592	– 2,179
	0,7	0,179	0,288	– 0,966	– 0,168	– 1,417	– 1,203	– 3,536
	1	0,168	0,301	– 1,353	– 0,31	– 1,902	– 1,688	– 4,747
Exakt auf Gesamtmenge + MMK auf Restmenge: Ohne Geburtsort	0,1	– 0,43	0,249	– 5,67	– 0,242	– 1,308	– 8,742	– 8,776
	0,3	– 0,577	0,181	– 6,21	– 0,411	– 2,053	– 9,459	– 10,284
	0,7	– 0,907	0,133	– 7,295	– 0,848	– 3,715	– 10,861	– 13,264
	1	– 1,19	0,009	– 8,104	– 1,221	– 4,945	– 11,884	– 15,624
Exakt auf Gesamtmenge + CLK auf Restmenge: Alle Merkmale	0,1	– 9,618	– 2,826	– 24,727	– 3,104	– 5,266	– 39,438	**– 103,316**
	0,3	– 9,607	– 2,766	– 24,769	– 3,23	– 5,765	– 39,465	**– 102,943**
	0,7	– 9,604	– 2,764	– 24,897	– 3,529	– 6,977	– 39,562	**– 102,305**
	1	– 9,633	– 2,748	– 25,031	– 3,781	– 8,07	– 39,674	**– 102,327**
Exakt auf Gesamtmenge + CLK auf Restmenge: Ohne Geburtsort	0,1	– 10,828	– 3,288	– 25,455	– 3,395	– 5,793	**– 40,533**	**– 107,78**
	0,3	– 10,765	– 3,206	– 25,605	– 3,616	– 7,017	**– 40,655**	**– 107,314**
	0,7	– 10,688	– 3,264	– 26,009	– 4,179	– 9,74	**– 40,971**	**– 107,278**
	1	– 10,672	– 3,292	– 26,396	– 4,674	– 11,998	**– 41,299**	**– 107,874**
Exakt auf Gesamtmenge + PRL auf Restmenge: Alle Merkmale	0,1	– 0,009	0,002	– 0,347	– 0,045	– 0,168	– 0,502	– 0,624
	0,3	– 0,133	– 0,036	– 0,703	– 0,134	– 0,431	– 0,94	– 1,755
	0,7	– 0,332	– 0,072	– 1,426	– 0,298	– 1,006	– 1,828	– 4,213
	1	– 0,494	– 0,118	– 1,968	– 0,454	– 1,503	– 2,493	– 6,038
Exakt auf Gesamtmenge + PRL auf Restmenge: Ohne Geburtsort	0,1	– 10,88	– 3,383	– 24,687	– 3,242	– 5,28	– 39,088	**– 100,111**
	0,3	– 11,261	– 3,609	– 24,935	– 3,556	– 6,253	– 39,334	**– 100,622**
	0,7	– 12,055	– 4,269	– 25,471	– 4,207	– 8,466	– 39,876	**– 101,941**
	1	– 12,691	– 4,73	– 25,895	– 4,732	– 10,195	**– 40,297**	**– 103,072**

Kritische Werte über 40 sind hervorgehoben.
SLK-581: Statistical Linkage Key; MMK: Multiple Matchkeys; CLK: Cryptographic Long-term Keys; PRL: Probabilistisches Record-Linkage

6

Diskussion und Fazit

Der Beitrag zeigt auf, wie die Mikrosimulation eines Registers genutzt werden kann, um Record-Linkage-Probleme bei der Erstellung und Verwendung eines Registers zu klären. Konkret wurde hierfür die Mikrosimulation eines bundesweiten Bildungsverlaufsregisters verwendet. Das simulierte Register wurde mittels verschiedener Record-Linkage-Verfahren verknüpft und diese Verfahren hinsichtlich ihrer Linkage-Qualität und eines möglichen Linkage-Bias untersucht.

Der Vergleich der Record-Linkage-Methoden zur Längsschnittverknüpfung des Bildungsverlaufsregisters zeigt, dass ein Linkage-Bias unter bestimmten Bedingungen auftritt. So weisen besonders Cryptographic Long-term Keys, deterministische Matchkeys und probabilistisches Record-Linkage (ohne Geburtsort) einen Linkage-Bias für die Gruppe der Migranten auf. Hiervon sind besonders weibliche Migranten, die während ihres Bildungsverlaufs heiraten, betroffen. Für die Umsetzung eines Bildungsverlaufsregisters bedeutet dies, dass insbesondere die Verknüpfung bei der Erwachsenenbildung Probleme verursachen kann, da für diese Gruppe die Wahrscheinlichkeit einer zwischenzeitlichen Heirat am größten ist.

Die besten Linkage-Ergebnisse und der geringste Linkage-Bias konnten durch probabilistisches Record-Linkage mit den Quasi-Identifikatoren Vorname, Nachname, Geburtsdatum, Geschlecht und Geburtsort erreicht werden. Dies entspricht den bisherigen Ergebnissen aus anderen Vergleichsstudien (Campbell, 2009; Gomatam und andere, 2002).

Auch Multiple Matchkeys wiesen sehr gute Ergebnisse auf, wobei dies darauf zurückzuführen ist, dass die verknüpften Datensätze jeweils exakte Teilmengen darstellen. Das Verfahren liefert deutlich schlechtere Ergebnisse, falls diese Bedingung nicht erfüllt ist (Weiand, 2024).

Die Ergebnisse zeigen, dass Linkage-Qualität und Linkage-Bias eng miteinander zusammenhängen. So weisen Verfahren mit hoher Linkage-Qualität einen niedrigen Linkage-Bias auf. Die Linkage-Qualität hängt dabei im Wesentlichen von der Datenqualität, Stabilität und Eindeutigkeit der Merkmale ab (Herzog und andere, 2007). Fehlertolerante Verfahren, wie Cryptographic Long-term Keys, Multiple-Matchkeys und probabilistisches Record-Linkage, zeigen dabei einen deutlich schwächeren Einfluss unterschiedlicher Datenqualität. Dies zeigt sich nicht nur beim Einfluss unterschiedlicher Datenqualitätsannahmen über das Register insgesamt, sondern auch bei den Unterschieden zwischen Einheimischen sowie Migrantinnen und Migranten. Deterministische Verfahren erzielen schlechtere Ergebnisse, ebenso Verfahren, die auf Verschlüsselungen basieren. Dies entspricht bisherigen Ergebnissen (Christen und andere, 2020). Solche Verfahren führen zu einem Linkage-Bias, nicht nur bei Migrantinnen und Migranten, sondern auch bei Frauen, da hier Namensänderungen bei Heirat häufiger auftreten.

LITERATURVERZEICHNIS

Campbell, Kevin M. *Impact of record-linkage methodology on performance indicators and multivariate relationships*. In: Journal of Substance Abuse Treatment. Jahrgang 36. Ausgabe 1/2009, Seite 110 ff. DOI: 10.1016/j.jsat.2008.05.004

Christen, Peter. *Data Matching. Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Berlin, Heidelberg 2012. DOI: 10.1007/978-3-642-31164-2

Christen, Peter/Ranbaduge, Thilina/Schnell, Rainer. *Linking Sensitive Data. Methods and Techniques for Practical Privacy-Preserving Information Sharing*. Cham 2020. DOI: 10.1007/978-3-030-59706-1

Collins, Linda M./Schafer, Joseph L./Kam, Chi-Ming. *A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures*. In: Psychological Methods. Jahrgang 6. Ausgabe 4/2001, Seite 330 ff.

Damerau, Fred J. *A Technique for Computer Detection and Correction of Spelling Errors*. In: Communications of the ACM. Jahrgang 7. Ausgabe 3/1964, Seite 171 ff. DOI: 10.1145/363958.363994

Fellegi, Ivan P./Sunter, Alan B. *A Theory for Record Linkage*. In: Journal of the American Statistical Association. Jahrgang 64. Nr. 328. 1969. Seite 1183 ff.

Gawronski, Katharina. *Konzeption eines Bildungsregisters in Deutschland*. In: WISTA Wirtschaft und Statistik. Ausgabe 2/2020, Seite 37 ff.

Giar, Katharina/Hohlstein, Franziska/Wipke, Mirco/Scharnagl, Alexander. *Konzeption eines Statistischen Bildungsverlaufsregisters in Deutschland – Entwicklungen bis 2023 und Ausgestaltungsoptionen*. In: WISTA Wirtschaft und Statistik. Ausgabe 3/2023, Seite 51 ff.

Gomatam, Shanti/Carter, Randy/Ariet, Mario/Mitchell, Glenn. *An empirical comparison of record linkage procedures*. In: Statistics in Medicine. Jahrgang 21. Ausgabe 10/2002, Seite 1485 ff. DOI: 10.1002/sim.1147

Haußmann, Michael. *Registermodernisierung und Zensus post-2021: Der Weg zu einem modernen amtlich-statistischen System in Deutschland*. In: Meinel, Gotthard/Schumacher, Ulrich/Behnisch, Martin/Krüger, Tobias (Herausgeber). Flächennutzungsmonitoring X. Flächenpolitik – Flächenmanagement – Indikatoren. IÖR Schriften Band 76, Seite 209 ff. Berlin 2018.

Herzog, Thomas N./Scheuren, Fristz J./Winkler, William E. *Data Quality and Record Linkage Techniques*. New York 2007.

Karmel, Rosemary. *Data linkage protocols using a statistical linkage key*. Australian Institute of Health and Welfare. Canberra 2005.

Kristensen, Thomas G./Nielsen, Jesper/Pedersen, Christian NS. *A tree-based method for the rapid screening of chemical fingerprints*. In: Algorithms for Molecular Biology Jahrgang 5. Artikel 9 (2010). DOI: 10.1186/1748-7188-5-9

LITERATURVERZEICHNIS

Kvalsvig, Amanda/Gibb, Sheree/Teng, Andrea. *Linkage error and linkage bias: A guide for IDI users*. Wellington 2019.

Li, Jinjing/O'Donoghue, Cathal. *A survey of dynamic microsimulation models: Uses, model structure and methodology*. In: International Journal of Microsimulation. Jahrgang 6. Ausgabe 2/2013, Seite 3 ff. DOI: 10.34196/ijm.00082

Peterson, James L. *A Note on Undetected Typing Errors*. In: Communications of the ACM. Jahrgang 29. Ausgabe 7/1986, Seite 633 ff. DOI: 10.1145/6138.6146

Randall, Sean/Brown, Adrian P./Ferrante, Anna M./Boyd, James H. *Privacy preserving linkage using multiple dynamic match keys*. In: International Journal of Population Data Science. Jahrgang 4. Ausgabe 1/2019. DOI: 10.23889/ijpds.v4i1.1094

RatSWD (Rat für Sozial- und WirtschaftsDaten). *Aufbau eines Bildungsverlaufsregisters: Datenschutzkonform und forschungsfreundlich.* RatSWD Positionspapier. [Zugriff am 17. Juni 2025]. Verfügbar unter: www.konsortswd.de

Schnell, Rainer. *Eignung von Personenmerkmalen als Datengrundlage zur Verknüpfung von Registerinformationen im Integrierten Registerzensus*. No. WP-GRLC-2019-01. German Record Linkage Center Working Paper Series Band 1/2019. DOI: 10.17185/DUEPUBLICO/49551

Schnell, Rainer. *Verknüpfung von Bildungsdaten in einem Bildungsregister mittels Record-Linkage auf Basis von Personenmerkmalen*. German Record Linkage Center Working Paper Series Band 3/2022. DOI: 10.17185/duepublico/76331

Schnell, Rainer/Bachteler, Tobias/Reiher, Jörg. *Privacy-preserving record linkage using Bloom filters*. In: BMC Medical Informatics and Decision Making. Jahrgang 9. Artikel 41/2009. DOI: 10.1186/1472-6947-9-41

Schnell, Rainer/Bachteler, Tobias/Reiher, Jörg. *A Novel Error-Tolerant Anonymous Linking Code*. German Record Linkage Center Working Paper Series Band 2/2011. DOI: 10.2139/ssrn.3549247

Schnell, Rainer/Weiand, Severin V. *Microsimulation of an educational attainment register to predict future record linkage quality*. In: International Journal of Population Data Science. Jahrgang 8. Ausgabe 1/2023. DOI: 10.23889/ijpds.v8i1.2122

Statistische Ämter des Bundes und der Länder. *Qualitätshandbuch der Statistischen Ämter des Bundes und der Länder (Version 1.21)*. Wiesbaden 2021.

Weiand, Severin V. *Vergleich von Record-Linkage Methoden anhand der Mikrosimulation eines bundesweiten Schülerregisters*. German Record Linkage Center Working Paper Series Band 4/2022. DOI: 10.17185/DUEPUBLICO/76361

Weiand, Severin V. *Analyse der Fehler in Quasi-Identifikatoren in einem deutschen Schülerregister durch probabilistische Längsschnittverknüpfung*. German Record Linkage Center Working Paper Series Band 2/2024. DOI: 10.17185/DUEPUBLICO/81929

Herausgeber
Statistisches Bundesamt (Destatis), Wiesbaden

Schriftleitung
Dr. Daniel Vorgrimler
Redaktion: Ellen Römer

Ihr Kontakt zu uns
www.destatis.de/kontakt

Erscheinungsfolge
zweimonatlich, erschienen im August 2025
Ältere Ausgaben finden Sie unter www.destatis.de sowie in der Statistischen Bibliothek.

Artikelnummer: 1010200-25004-4, ISSN 1619-2907