German Record Linkage Center
WORKING PAPER SERIES
NO. WP-GRLC-2024-02 | MAY 6, 2024

Analyse der Fehler in Quasi-Identifikatoren in einem deutschen Schülerregister durch probabilistische Längsschnittverknüpfung

Severin Vincent Weiand

Inhaltsverzeichnis

Das vorliegende Working-Paper wurde 2024 als Masterarbeit im Studiengang Survey-Methodology an der Universität Duisburg-Essen am Lehrstuhl für Empirische Sozialforschung erstellt.

Zusammenfassung

Zur Verbesserung des deutschen Schulsystems setzen sich zurzeit Bund und Länder mit den Fragen zur Ausgestaltung eines bundesweiten Bildungsverlaufsregisters auseinander. Dieses Register soll die Analyse von Bildungsverläufen im Längsschnitt ermöglichen und mit Record-Linkage erstellt werden. Die Verknüpfung soll anhand von Quasi-Identifikatoren (QIDs), wie Name oder Geburtsdatum, erfolgen. Die Implementation des Registers benötigt daher die Spezifikation über die benötigten QIDs. Für diese Spezifikation bedarf es genauer Informationen über die zu erwartende Qualität der Daten in einem Bildungsverlaufsregister sowie einer Abschätzung der erreichbaren Linkage-Qualität und eines potenziellen Linkage-Bias. Diese Arbeit führt hierzu eine Datenqualitätsanalyse sowie eine Verknüpfung des Zentralen Schülerregisters (ZSR) der Stadt Hamburg durch. Das ZSR enthält einen eindeutigen Identifikator, anhand dessen die Ergebnisse überprüft werden (Goldstandard-Datensatz).

Die beste Linkage-Qualität konnte durch probabilistisches Record-Linkage erreicht werden. Es zeigt sich, dass für ein unverzerrtes Verknüpfungsergebnis eine Adressangabe benötigt wird. Hierzu eignen sich Gitterzellen-Koordinaten nach dem INSPIRE-Referenzsystem. Die Staatsangehörigkeit erweist sich hingegen als kein für die Verknüpfung relevantes Merkmal. Trotz der hohen erreichbaren Verknüpfungsrate wird eine nachträgliche manuelle Klassifikation von nicht eindeutigen Record-Paaren notwendig, da ansonsten zahlreiche Bildungsverläufe unvollständig oder falsch abgebildet werden. Diese falschen Verknüpfungen ohne manuelle Klassifikation betreffen besonders Migranten und Schüler von Stadtteilschulen.

Abstract

The German federal and state administrations are discussing the design of a nationwide educational history register. This register should enable the longitudinal analysis of educational careers. A longitudinal register requires techniques for linking information on the same person over time using quasi-identifiers (QIDs) such as name or date of birth. Therefore, indications about the necessary QIDs are needed to implement the register. For this, further information about the expected data quality in an education history register and the expected linkage quality and linkage bias are required. To this end, this paper conducts a data quality analysis and a record linkage of the pupil register of the City of Hamburg (Zentrales Schülerregister, ZSR) using an administrative unique identifier to validate the linkage.

The best linkage quality was achieved by probabilistic record linkage. Additional information concerning addresses is required to prevent subgroups from differing in their probability of successful linkage (linkage bias). Grid cell coordinates according to the INSPIRE reference system suit this purpose. On the other hand, the linkage results indicate that citizenship is not a relevant feature for the linkage. A manual classification of record pairs is nevertheless necessary. Otherwise, some education histories created by the linkage will be incomplete or incorrect. The results show that migrants and pupils from district schools (a peculiarity in Hamburg) are particularly affected by this linkage bias.

Schlagwörter: Record-Linkage, Bildungsverlaufsregister, Datenqualität, Registerzensus, Linkage-Bias, data quality, educational register

Danksagung: Ich danke Herrn Prof. Dr. Rainer Schnell und Herrn Dr. habil. Tobias Brändle für die Idee, die Initiierung und die Betreuung dieser Arbeit.

1. Einleitung

Zurzeit setzen sich Bund und Länder mit Fragen zur Ausgestaltung eines bundesweiten Bildungsverlaufsregisters auseinander.[1] Das Register soll über Record-Linkage aus einer Vielzahl von Datenquellen zusammengestellt werden. Zur Erstellung eines solchen Registers benötigt es ein funktionierendes Identitätsmanagement, um Personen über Landesgrenzen und Bildungsbereiche hinweg zu verknüpfen. Bevor das Register implementiert werden kann, bedarf es einer Entscheidung über erforderliche Personenmerkmalen, die für die Verknüpfung und somit ein funktionierendes Identitätsmanagement benötigt werden. Hierbei muss eine Balance zwischen Datensparsamkeit und Datenqualität gefunden werden. Ein Gutachten für das Bundesministerium für Bildung und Forschung (BMBF) auf Basis einer Mikrosimulation des Bildungsverlaufsregisters sowie weiterer darauf aufbauender Publikationen wurden hierzu bereits veröffentlicht (Schnell 2022; Weiand 2022; Schnell & Weiand 2023).

In der Mikrosimulation wurden die Personenmerkmale des Bildungsverlaufsregisters zunächst generiert und danach Änderungen im Register simuliert (z. B. Umzüge, Schulabschlüsse oder neue Schülerkohorten). Generierung und Simulation erfolgten anhand von bekannten Verteilungen sowie Annahmen über die Datenqualität der Personenmerkmale. Insbesondere über die zu erwartende Datenqualität von Personenmerkmalen in deutschen administrativen Datensätzen gab und gibt es keine Informationen. Dies stellt jedoch ein zentrales Problem dar, denn die Datenqualität ist der wesentliche Faktor, um Aussagen über die Belastbarkeit der Simulationsergebnisse treffen zu können.

In der folgenden Arbeit wird daher eine Analyse der Daten- und Verknüpfungsqualität eines deutschen administrativen Datensatzes vorgestellt. Datengrundlage ist das Zentrale Schülerregister (ZSR) sowie der Schülerindividualdatensatz (IVDS) der Freien und Hansestadt Hamburg für die Schuljahre 2021/22 und 2023/24. Das ZSR ist das einzige landesweite Schülerregister in Deutschland und verfügt über ein bereits implementiertes Identitätsmanagement. Der im ZSR geführte eindeutige Identifikator liegt für alle gelieferten Datensätze vor. Dieser Identifikator ermöglicht sowohl die Evaluation der Datenqualität des ZSR als auch die Qualität von Record-Linkage-Verfahren mit verschiedenen Personenmerkmalen. Die Ergebnisse sollen sowohl wichtige Informationen über die Belastbarkeit der Simulationsergebnisse sowie weiteren Aufschluss über die Ausgestaltung des bundesweiten Bildungsverlaufsregisters liefern.

Folgende Fragen sollen hierzu beantwortet werden:

1. Welche personenbezogenen Merkmale werden benötigt, um einen möglichst hohen Verknüpfungserfolg von bildungsbereichs- und länderübergreifenden Bildungsdaten zu erzielen?
2. Welcher Verknüpfungserfolg ist zu erwarten, wenn ausschließlich unveränderliche Merkmale aus den Bildungsstatistiken für die Verknüpfung genutzt werden?
3. Welcher Verknüpfungserfolg ist zu erwarten, wenn die ausgewählten Merkmale vor der Verknüpfung mittels Verschlüsselung in eine ID überführt werden und eine Re-Identifikation nicht möglich ist (z. B. durch Bildung eines nicht umkehrbaren Hash-Codes aus den ausgewählten Merkmalen)?

[1] In der aktuellen Diskussion wird das Bildungsverlaufsregister auch als Schülerregister oder Bildungsregister bezeichnet. Die Bezeichnungen sind größtenteils synonym.

4. Ist es möglich, ein fehlertolerantes Verschlüsselungsverfahren zur ID-Generierung einzusetzen? Welches Verfahren wäre denkbar und wie wirkt sich das auf den Verknüpfungserfolg aus?
5. Wie groß ist die Stabilität der zur Verknüpfung verwendeten Merkmale über die Zeit?
6. Werden durch die Verknüpfung in einem Bildungsregister bestimmte Subgruppen (z. B. Personen mit Migrationsstatus) aufgrund der Instabilität der Merkmale besonders häufig aus den Analysen der Bildungsverläufe ausgeschlossen?
7. Welche Fehlerquoten und welche Art von Fehlern sind für die einzelnen Merkmale zu erwarten?

Zur Beantwortung der Fragen ist die Arbeit wie folgt strukturiert: In Abschnitt 2 wird die Problemstellung weiter präzisiert. Dazu werden die aktuelle und historische Diskussion über ein bundesweites Bildungsverlaufsregister sowie die Anforderungen und Erwartungen an das Register zusammengefasst. Abschnitt 3 befasst sich mit notwendigen theoretischen Grundlagen zu Record-Linkage und Datenqualität. Hierbei werden relevante Einflussfaktoren für die Verknüpfung ausgearbeitet, welche im weiteren Verlauf analysiert werden müssen. Abschnitt 4 befasst sich mit der Deskription der Daten, insbesondere der Datenqualität und der Beschreibung der Fehlerarten. Zudem erfolgt eine detaillierte Beschreibung der Datenquellen. In Abschnitt 5 wird darauf aufbauend die Verknüpfung der Daten anhand mehrerer Record-Linkage-Verfahren und die Analyse der Record-Linkage-Ergebnisse beschrieben.

2. Präzisierung der Problemstellung

Zur Präzisierung des Untersuchungsgegenstands wird im Folgenden die Diskussion über ein bundesweites Bildungsverlaufsregister zusammengefasst. Erwarteter Nutzen des Registers sowie bereits angemerkte mögliche Probleme stehen dabei im Fokus.

2.1. Aktuelle und historische Diskussion zum Bildungsverlaufsregister

Ein bundesweites Bildungsverlaufsregister wird in Deutschland schon lange diskutiert. Bereits im Januar 2000 plante die Kultusministerkonferenz (KMK) eine Vereinheitlichung der Bildungsdaten auf Landesebene (Brameshuber 2007). Hieraus folgte 2003 die Ausarbeitung eines Kerndatensatzes mit Individualdaten, der von den Ländern erhoben werden sollen (Kultusministerkonferenz 2004). Obwohl die Empfehlung 2003 ausgesprochen und von den Kultusministern einstimmig gefasst wurde, ist der Beschluss zum Kerndatensatz bis heute nicht vollständig umgesetzt (Mundelius 2019; Hertweck et al. 2023).

Nach dem Beschluss der Kultusministerkonferenz stagnierte die Diskussion um ein bundesweites Bildungsverlaufsregister jedoch größtenteils. Ursache hierfür waren besonders datenschutzrechtliche Bedenken, die auch zur Verleihung des Big Brother Awards im Jahr 2006 führten.[1] Dies sorgte zunächst für ein Ende der Diskussion über ein bundesweites Bildungsverlaufsregister. Erst 2018 wurde die Diskussion durch die Pläne eines Registerzensus wiederbelebt (Mundelius 2019). Das ursprünglich alleinige Ziel der Untersuchung individueller Bildungsverläufe wurde durch Lieferverpflichtungen eines Zensus gegenüber der EU erweitert. Insbesondere muss der Bildungsstand aller in Deutschland wohnhaften Personen über 15 Jahre erfasst werden (Gawronski 2020). Die Anforderungen an einen Zensus könnten durch eine Erfassung der Bildung im Querschnitt, in Form eines Zensusbildungsregisters, bereits erfüllt werden. Aufgrund der zahlreichen analytischen Vorteile wird jedoch bisher ein Bildungsverlaufsregister einem Zensusbildungsregister vorgezogen (Gawronski 2020). Für ein solches Vorgehen sprach sich auch der RatSWD (2023) aus.

Zentral zur Erfüllung der Anforderungen an ein Bildungsverlaufsregister ist ein funktionierendes Identitätsmanagement. Dabei erhalten alle im Register enthaltenen Personen eine eindeutige Identifikationsnummer (ID), mit der ihre Bildungsverläufe über die Jahre analysiert werden können. Diese Identifikationsnummer wird in der Diskussion zumeist als Bildungs-ID oder Schüler-ID bezeichnet, wobei die Begriffe synonym sind.

Für das Identitätsmanagement des Registers ist eine eigens eingerichtete unabhängige dritte Stelle geplant. Diese wird in der Diskussion meist als Trusted Third Party oder Vertrauensstelle bezeichnet (Giar et al. 2023; Gawronski 2020). In der Internationalen-Literatur wird eine solche Stelle als *Linkage-Unit* bezeichnet (Christen et al. 2020: 82).

[1] https://bigbrotherawards.de/2006/kultusministerkonferenz (abgerufen am 22.04.2024)

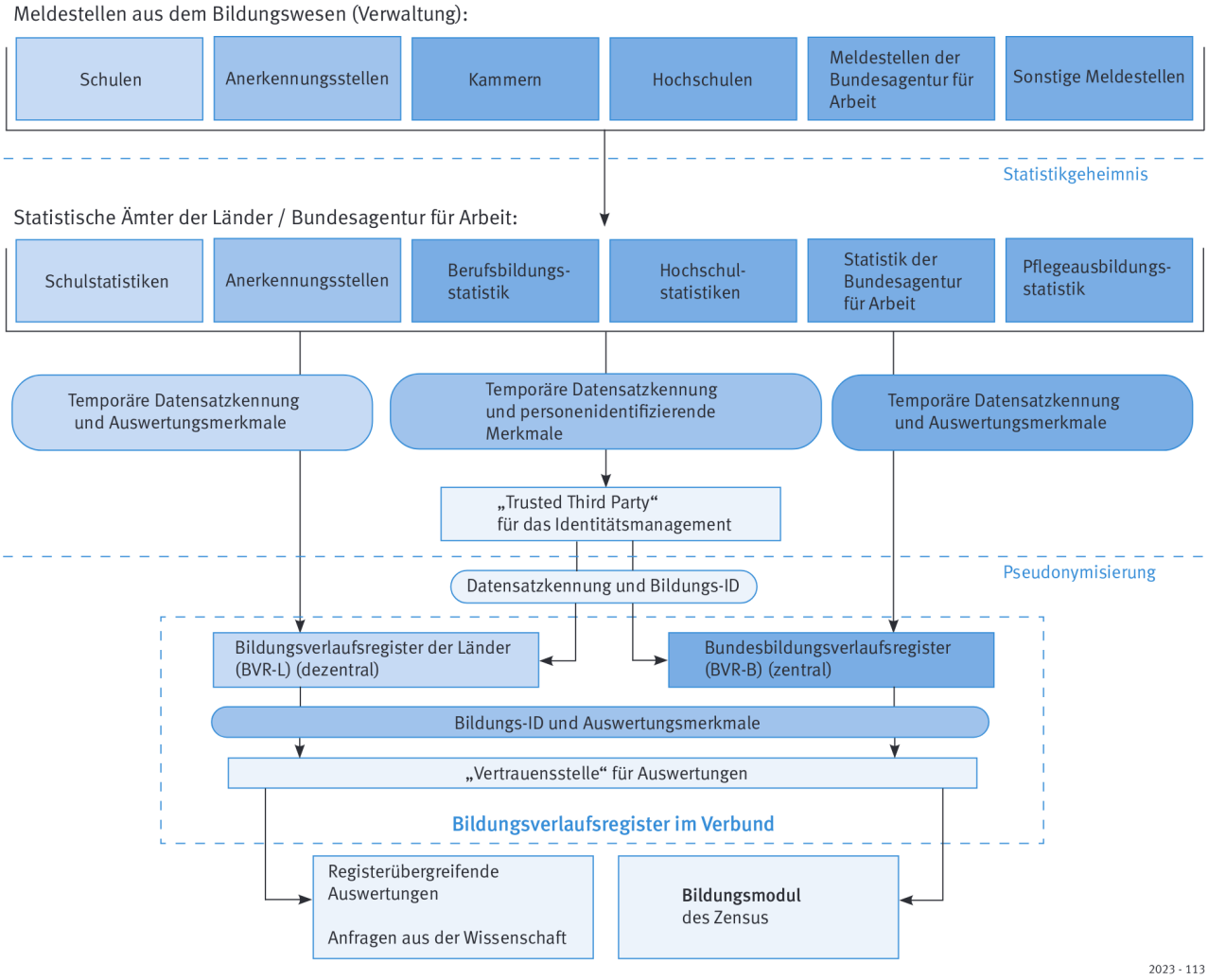


Abbildung 2.1.: Datenflussdiagramm für das Bildungsverlaufsregister der Länder (BVR-L) und des Bundes (BVR-B) (Giar et al. 2023: 55).

Die aktuelle Planung sieht vor, dass die Daten von Personen im Bildungswesen zunächst dezentral von den Ländern gesammelt und zusammengeführt werden. Abbildung 2.1 stellt hierzu ein von Giar et al. (2023) erstelltes Datenflussdiagramm dar. Zur Erstellung des bundesweiten Registers soll demnach ein Satz an Personenmerkmalen sowie die entsprechenden Datensatz-IDs an die Linkage-Unit weitergegeben werden. Die Linkage-Unit verknüpft dann mittels Record-Linkage die Personen mit dem bestehenden Register. Kann eine Person mit einer vorhandenen Person verknüpft werden, so erhält sie die Bildungs-ID der bereits bekannten Person, andernfalls erhält sie eine neue, bislang nicht vergebene Bildungs-ID.

Die Kombination aus Datensatz-ID und Bildungs-ID wird Bund und Ländern übermittelt. Bund und Ländern können daraus sowohl die dezentralen Bildungsverlaufsregister der Länder (BVR-L) als auch das zentrale Bildungsverlaufsregister des Bundes (BVR-B) erstellen.

Das Datenflussdiagramm zeigt, dass eine Vielzahl von unterschiedlichen Meldestellen zur Erstellung des Bildungsverlaufsregisters beitragen. Dies kann zu einigen Problemen führen, da ggf. Zuständigkeiten für die Erfassung einzelner Personen unklar sind. Dadurch können Under- und Overcoverage entstehen (Schnell 2019a). Dies bedeutet, dass sowohl Personen erfasst werden können, die nicht durch

das Register erfasst werden sollen (Overcoverage), als auch Personen nicht erfasst werden, die jedoch durch das Register erfasst werden sollten (Undercoverage). Weiter können unterschiedliche Methoden und Vorgaben der Erfassung zwischen den Meldestellen und Bundesländern zu späteren Problemen bei der Verknüpfung führen (Schnell 2019a; Christen & Schnell 2023). Da in dieser Arbeit nur die Daten der Hamburger Schulen zur Verfügung stehen, kann das Ausmaß dieser Probleme nicht geprüft werden.

2.2. Anforderungen an ein Bildungsverlaufsregister

Ein Bildungsverlaufsregister kann zahlreiche Möglichkeiten zur Analyse des Schulsystems und Bildungserfolgs bieten, die aus Kostengründen für Surveys kaum zu leisten sind. In der Literatur finden sich einige Erwartungen, die mit dem Register verknüpft werden, und im Folgenden zusammengefasst werden (Hertweck et al. 2023; RatSWD 2023; Gawronski 2020; Giar et al. 2023; Baas 2021). Im Rahmen dieser Arbeit gilt es dabei zu prüfen, ob ein Bildungsverlaufsregister in der aktuellen Planung diesen Erwartungen gerecht werden kann.

Eine zentrale Erwartung an das Register ist die Analyse von Bildungsverläufen, die Schulwechsel oder Unterbrechungen vorweisen. Auch die Möglichkeit zur Analyse des Bildungserfolgs und der Bildungsbeteiligung bestimmter Subpopulationen wird als konkrete Erwartung an das Register gestellt. Zu diesen Subpopulationen zählen insbesondere Migranten und Menschen mit Migrationshintergrund. Weiter wird auch erwartet, den Übergang zwischen Schule und Hochschule über Bundesländer hinweg analysieren zu können. Ferner sollen aufgrund der Vollerhebung kleinräumige Ergebnisse, wie etwa auf Gemeinde- oder Stadtteilebene, analysiert werden können. Im Zeitverlauf können zudem die Auswirkungen politischer Beschlüsse und anderer Ereignisse analysiert werden. Erwartet wird somit, dass ein Bildungsverlaufsregister eine einzigartige Datenquelle zur Analyse und Verbesserung des Schulsystems, insbesondere zur Erreichung von Bildungsgerechtigkeit darstellt.

Von wesentlicher Bedeutung zur Erfüllung dieser Anforderungen ist die Qualität der Verknüpfung. Für den Erfolg der Verknüpfung ist wiederum die Art, Anzahl und Qualität der zu übermittelnden Merkmale von zentraler Bedeutung. Eine Expertise zu dem Thema auf Literaturbasis wurde von Schnell (2019a) erstellt. Um die Realisierbarkeit des Registers weiter zu untersuchen, wurde in dieser Expertise eine Simulation empfohlen, welche auch durchgeführt wurde (Schnell 2022; Weiand 2022; Schnell & Weiand 2023). Besonders anhand der in diesen Untersuchungen hervorgegangenen Probleme und Unklarheiten sollen die Ziele dieser jetzigen Arbeit noch einmal konkretisiert und präzisiert werden.

Zunächst konnte festgestellt werden, dass keine ausreichenden Informationen über die zu erwartende Datenqualität des Registers zur Verfügung stehen. Dies bedeutet, dass unklar ist, wie viele Fehler (Tippfehler, unvollständige Angaben) bei einzelnen Merkmalen zu erwarten sind. Aus diesem Grund wurden in der Simulation vier verschiedene Fehlerquoten getestet (0.1 %, 0.3 %, 0.7 % und 1 %), die jeweils für alle Merkmale konstant waren. Veränderungen der Daten, welche einen Fehler auslösen, wurden durch verschiedene Ereignisse simuliert (Umzug, Bildungsübergänge, Heirat). Es wurde zudem angenommen, dass keine fehlenden Werte vorliegen. Die Probleme zeigen, dass eine empirische Betrachtung der Datenqualität zur Einordnung der Ergebnisse der Simulation von großer Bedeutung sind. Dies soll daher im Rahmen dieser Arbeit erfolgen.

Die Mikrosimulation erstellte jedes Jahr einen neuen Registerdatensatz und verknüpfte diesen mit dem bestehenden Datensatz. Wie bei den aktuellen Plänen zum Bildungsverlaufsregister vorgesehen, wurden in der Simulation keine Fälle gelöscht. Das Register akkumuliert somit jährlich immer mehr Personen. Aufgrund von Fehlern und Personen mit ähnlichen oder gleichen Merkmalsausprägungen

führte das Anwachsen des Registers zu immer höheren Wahrscheinlichkeiten für Fehlverknüpfungen. Da Personen im ZSR nicht dauerhaft gespeichert werden (siehe Abschnitt 4.1) und das ZSR eine deutlich geringe Anzahl Personen erfasst, können diese Befunde im Rahmen dieser Arbeit nicht geprüft werden. Zudem ist die Wahrscheinlichkeit für Personen mit gleichen Merkmalsausprägungen für ein einzelnes Bundesland wie Hamburg deutlich geringer als in einem bundesweiten Register. Dennoch können Probleme, die bereits bei einer kleinen Datenbasis auftreten, einen großen Aufschluss über zukünftige Probleme liefern.

Von wesentlicher Bedeutung für die derzeitige Diskussion ist die Anzahl und Art der Merkmale, die für das Record-Linkage zur Verfügung stehen. Dabei steht die Anforderung der Datensparsamkeit, möglichst wenige Merkmale zu verwenden, der Linkage-Qualität gegenüber, die sich bei der Einbeziehung weiterer Merkmalen zumeist verbessert (siehe Abschnitt 3.1). Das Ergebnis der Simulation war, dass Vorname, Nachname, Geburtsdatum (Tag, Monat, Jahr), Geschlecht und Geburtsort mindestens notwendig sind, um eine erfolgreiche Verknüpfung auf Bundesebene gewährleisten zu können.

Weitere in der Diskussion zum Bildungsverlaufsregister genannten Merkmale zur Verknüpfung sind (Schnell 2019a; Gawronski 2020; Schnell 2022): Adressdaten (Straße, Hausnummer, PLZ, Stadt, Bundesland), Staatsangehörigkeit, Geburtsland, Familienstand sowie Angaben zu Ehepartnern, Kindern und zur Schule (Adresse, Name). Nicht alle der genannten Merkmale finden sich in den zur Verfügung stehenden Daten und können daher getestet werden. Dennoch besteht das Ziel dieser Arbeit darin, weiteren Aufschluss über die notwendigen Merkmale zu liefern.

Als Merkmal bisher nicht diskutiert wurde eine Gitterzellen-Koordinate nach dem INSPIRE-Referenzsystem (Infrastructure for Spatial Information in Europe). Eine Gitterzellen-Koordinate würde dabei einige Vorteile gegenüber einer herkömmlichen Adresse bringen. So würde aus datenschutzrechtlicher Perspektive keine genaue Wohnadresse gespeichert und für die Verknüpfung übermittelt werden, sondern die ID der Gitterzellen, in der sich die wahre Adresse befindet. INSPIRE ist europaweit einheitlich für eine Vielzahl von Gittergrößen definiert, die hierzu verwendet werden können (INSPIRE 2014). Die Ergebnisse des Zensus 2011 wurden z. B. in den Gittergrößen 100x100 m und 1x1 km veröffentlicht.[1]

Auch für die Verknüpfung ergeben sich Vorteile durch die Verwendung einer Gitterzellen-Koordinate. Im Gegensatz zu anderen Adressmerkmalen ändert sich die Gitterzellen-ID nicht durch Umbenennung oder Grenzverschiebung. Zudem kann eine approximative Distanz zwischen zwei Punkten berechnet werden. So kann für die Verknüpfung die geografische Distanz zweier Adressen einbezogen werden, da Menschen häufig über kurze Distanzen umziehen (siehe Abschnitt 4.4.2). Für eine Analyse kann die Adresse genutzt werden, um beispielsweise die Distanz zwischen Wohnort und Schule zu berechnen.

Die Auswirkung auf die Linkage-Qualität durch die Verwendung einer Gitterzellen-ID soll im Rahmen dieser Arbeit untersucht werden. Es besteht sowohl die Möglichkeit, dass sich die Qualität durch die Ungenauigkeit verbessert als auch dass sie sich verschlechtert. Wesentlicher Faktor ist dabei die verwendete Gittergröße.

Als Letztes besteht die Frage zur verwendeten Record-Linkage-Methode und der zu verwendenden Parameter. Folgende Methoden wurden in der ersten Expertise veranschlagen und in der Simulation sowie in einer weiteren Untersuchung getestet[2] (Schnell 2019a; Schnell 2022; Weiand 2022):

1. Matchkeys,
2. Cryptographic Longterm Keys (CLKs),

[1] https://www.zensus2011.de/DE/Home/Aktuelles/DemografischeGrunddaten.html (abgerufen am 02.03.2023)

[2] Nicht genannt ist die Signature-Methode. Die bisherigen Untersuchungen zeigten, dass die Methode bislang nicht vollständig ausgereift ist, weshalb sie im nicht weiter betrachtet wurde.

3. probabilistisches Record-Linkage mittels ECM und
4. Multiple-Matchkeys.

Die bisherigen Ergebnisse zeigen, dass probabilistisches Record-Linkage die vielversprechendsten Ergebnisse liefert (Weiand 2022). Eine wesentliche Frage besteht zudem darin, ob die Verknüpfung des bundesweiten Registers verschlüsselt erfolgen kann (Schnell 2019a). Eine verschlüsselte Verknüpfung könnte dabei einige datenschutzrechtlichen Probleme lösen. Mehrere der genannten Methoden sind hierfür ausgelegt bzw. können entsprechend modifiziert werden.[1] Zum Vergleich der Ergebnisse werden alle genannten Methoden an den zur Verfügung stehenden Daten getestet. Sofern eine Methode einen Parameter zur Klassifikation verwendet, werden zudem unterschiedliche Einstellungen dieses Parameters getestet. Eine detaillierte technische Beschreibung hierzu erfolgt in Abschnitt 5.1.

[1] Matchkeys und Multiple-Matchkeys können durch die Verwendung einer kryptografischen Hashfunktion verschlüsselt werden. CLKs sind bereits verschlüsselt. Probabilistisches Record-Linkage kann durch die Verwendung einzelner Bloom-Filter auch verschlüsselt werden. Durch die getrennte Abspeicherung der Merkmale wäre eine solche Verschlüsselung jedoch angreifbar (Christen et al. 2020: 111).

3. Theoretische Grundlagen

3.1. Einführung in Record-Linkage

Als Record-Linkage wird die Verknüpfung der Informationen über dieselbe Person aus mehreren Datensätzen verstanden (Christen et al. 2020: 3). Die Verknüpfung kann sowohl im Querschnitt als auch im Längsschnitt erfolgen. Bei einer Querschnittsverknüpfung werden zwei oder mehr Datensätze aus verschiedenen Quellen verbunden, die zu einem ähnlichen Zeitpunkt erhoben wurden. Die Datensätze enthalten dabei zumeist verschiedene Variablen. Beispiel hierfür ist die Verknüpfung einer Lernstanderhebung von Schülern mit einem Schülerregister. Bei einer Längsschnittsverknüpfung werden Datensätze aus der gleichen oder unterschiedlichen Quellen verbunden, die zu verschiedenen Zeitpunkten erhoben wurden. Die Datensätze enthalten dabei zumeist die gleichen Variablen, die an unterschiedlichen Zeitpunkten gemessen wurden. Beispiel hierfür ist die Verknüpfung von Leistungsdaten einzelner Schüler über mehrere Jahre, um die Entwicklung der Schüler zu erfassen.

Der einfachste Fall zur Verknüpfung von Datensätzen liegt vor, wenn alle Datensätze über die gleiche eindeutige ID verfügen (Schnell 2016). Die Verknüpfung erfolgt dann anhand dieser Nummer. Aufgrund der Trivialität dieses Verfahrens wird eine solche Verknüpfung zumeist als *File-Merge* bezeichnet und wird im Folgenden nicht weiter behandelt.

Technisch schwieriger sind all jene Fälle, in denen keine eindeutige ID vorliegt. Die Verknüpfung kann in diesen Fällen nur anhand von *Quasi-Identifikatoren* (QIDs) erfolgen. Dies sind Merkmale, die eine Person nicht eindeutig identifizieren, wie z. B. der Name oder das Geburtsdatum. Nur durch eine Kombination mehrerer QIDs ist es möglich, die gleiche Person in beiden Datensätzen zu identifizieren. Je mehr Merkmale zur Verfügung stehen, desto weniger Fehler entstehen üblicherweise bei der Verknüpfung und desto eindeutiger können Personen identifiziert werden (Christen et al. 2020: 7 f.).

Zur Verknüpfung von Entitäten anhand von QIDs wurden eine Vielzahl von Record-Linkage-Methoden entwickelt.[1] Diese können in deterministische und probabilistische Methoden unterschieden werden (Herzog et al. 2007: 81). Bei einer deterministischen Methode werden Records nur miteinander verknüpft, falls eine vorher definierte Menge an QIDs exakt übereinstimmen. Eine Menge an QIDs, die für einen exakten Abgleich verwendet wird, wird zumeist als *Matchkey* bezeichnet. Eine probabilistische Methode berechnet aus der Menge an QIDs eine Wahrscheinlichkeit, dass beide Records zur gleichen Entität gehören. Zwei Records werden dann zumeist als Match klassifiziert, wenn diese Wahrscheinlichkeit über einem vorher festgelegten Schwellenwert liegt (Christen 2012: 133 f.). Die überwiegende Mehrheit der Literatur zeigt, dass probabilistische Methoden bessere Verknüpfungsergebnisse erzielen (Campbell 2009; Gomatam et al. 2002; Tromp et al. 2011; Bohensky 2016; Weiand 2022).

Eine weitere Unterscheidung besteht darin, ob eine Methode auch verschlüsselt durchgeführt werden kann, sodass die Partei, welche die Verknüpfung durchführt, die Identität der Personen nicht feststellen kann. Diese Methoden werden unter dem Oberbegriff Privacy Preserving Record-Linkage

[1] Im Folgenden wird als Record-Linkage-Methoden das theoretische Vorgehen und als Record-Linkage-Verfahren die konkrete praktische Anwendung zur Verknüpfung bezeichnet. Bei einem Record-Linkage-Verfahren stehen die Parameter für die Verknüpfung, wie z. B. Schwellenwerte und Blocking-Regeln, fest.

(PPRL) zusammengefasst. PPRL-Methoden können sowohl deterministisch als auch probabilistisch sein (Christen et al. 2020).

3.2. Gütemaße zur Bewertung der Linkage-Qualität

Zum Vergleich von Record-Linkage-Methoden benötigt es Angabe über die Qualität der Ergebnisse. Diese wird als *Linkage-Qualität* bezeichnet und kann anhand mehrerer im Folgenden dargestellter Gütemaße beurteilt werden. Grundlage für die Gütemaße sind die absoluten Zahlen der erfolgten richtigen und falschen Klassifikationen aller möglichen Record-Paare. Das Ergebnis einer Klassifikation hängt dabei vom Klassifizierten und wahren Zustand (true match state) eines Paars ab. Diese Dimensionen können in Form einer Kreuztabelle (Konfusionsmatrix) dargestellt werden (Abbildung 3.1; Christen 2012: 166). Korrekt verknüpfte Paare werden als *True Positive* (TP) und korrekt nicht verknüpfte Paare als *True Negative* (TN) bezeichnet. Werden zwei Records verknüpft, obwohl sie nicht zur gleichen Person gehören, handelt es sich um ein *False Positive* (FP). Bleiben zwei Records nicht verknüpft, obwohl sie zur gleichen Person gehören, so handelt es sich um ein *False Negative* (FN).

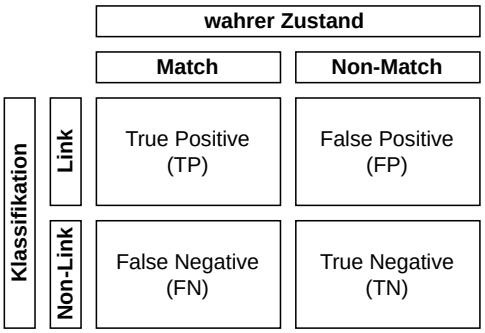


Abbildung 3.1.: Konfusionsmatrix

Die wichtigsten Gütemaße für Record-Linkage sind Precision und Recall. Sie sind definiert als (Christen 2012: 167):

$$Precision = \frac{TP}{TP + FP} \tag{3.1}$$

und

$$Recall = \frac{TP}{TP + FN}. \tag{3.2}$$

Recall ist auch als Sensitivität bekannt, Precision als Positiv Predictive Value. Precision und Recall stellen zwei Kriterien zur Auswahl eines Verfahrens dar. In der amtlichen Statistik wird zumeist ein Fokus auf Precision gelegt (Schnell 2019a). Entsprechend gilt es False Positives möglichst zu verhindern (FP-Null-Regel), auch wenn dies zu einer Reduktion des Recalls führt. Wie Schnell (2019a) darlegt, ist das strikte Einhalten der FP-Null-Regel für eine Verknüpfung nicht immer ratsam. Die in dieser Arbeit implementierten Record-Linkage-Verfahren wurden daher auf eine möglichst hohe Precision ausgelegt, kleinere Verschlechterungen der Precision wurden jedoch akzeptiert, falls eine deutliche Verbesserung des Recalls erreicht werden konnte.

Für die Optimierung eines Verfahrens auf Precision und Recall benötigt es zumeist eine Zusammenfassung der beiden Gütemaße. Dies erfolgt häufig in Form des F-Werts. Aufgrund zahlreicher Probleme des F-Werts bei Record-Linkage-Ergebnisse (Hand & Christen 2018) wird im Rahmen dieser Arbeit

das Gütemaß F^* verwendet (Hand et al. 2021):

$$F^* = \frac{TP}{TP + FP + FN}. \tag{3.3}$$

Ein wesentliches Problem der Verwendung dieser Gütemaße besteht darin, dass der True Match State eines Record-Paars bekannt sein muss. Datensätze, bei denen der True Match State bekannt ist, werden als *Goldstandard-Daten* bezeichnet (Harron et al. 2017; Christen 2012: 163). Goldstandard-Daten können durch eine Simulation oder durch Record-Linkage selbst erzeugt werden. Die erste Methode wurde bei der Mikrosimulation des Bildungsverlaufsregisters angewendet, die zweite Methode wird in dieser Arbeit angewendet.

Bei einer Simulation werden alle Records im Datensatz nach empirischen Verteilungen generiert. Der Vorteil besteht darin, dass der True Match State immer korrekt ist. Zudem kann jeder beliebige Maßstab getestet und jede QID erzeugt werden. Der Nachteil besteht darin, dass selten alle empirischen Verteilungen gegeben sind, sodass sich eine Simulation meistens von einem empirischen Datensatz unterscheidet (Christen 2012: 178 ff.).

Goldstandard-Daten können mittels Record-Linkage erstellt werden, indem die Daten durch ein anderes, besseres Record-Linkage-Verfahren verknüpft und idealerweise manuell validiert werden. Eine manuelle Klassifikation von ausgewählten Fällen wird als *Clerical-Review* bezeichnet (Christen 2012: 174). Der Vorteil bei der Verwendung von Record-Linkage zur Generierung von Goldstandard-Daten besteht darin, dass die Verknüpfung mit echten empirischen Daten getestet werden kann. Nachteile sind, dass der erzeugte True Match State Fehler enthalten kann und die Erzeugung eines solchen Datensatzes meist mit einem sehr hohen Aufwand verbunden ist (Christen 2012: 34 f.).[1]

3.3. Datenqualität

Im Kontext dieser Arbeit existieren zwei Betrachtungen von Datenqualität. Diese können darin unterschieden werden, ob das bundesweite Bildungsverlaufsregister oder das ZSR betrachtet wird. Wird das bundesweite Bildungsverlaufsregister – das verknüpfte Endprodukt – betrachtet, so besteht der Datenqualitätsanspruch an die Erfüllung der Anforderungen eines statistischen Produkts der amtlichen Statistik (Statistische Ämter des Bundes und der Länder 2021). Von den Qualitätskriterien, auf die sich Bund und Länder geeinigt haben, können anhand der vorliegenden Daten zwei Kriterien beurteilt werden:

1. Genauigkeit und Zuverlässigkeit: “Die Statistiken spiegeln die Realität genau und zuverlässig wider” (Statistische Ämter des Bundes und der Länder 2021: 12).
2. Kohärenz und Vergleichbarkeit: “Die Statistiken sind untereinander und im Zeitablauf konsistent und zwischen Regionen und Ländern vergleichbar; es ist möglich, miteinander in Beziehung stehende Daten aus unterschiedlichen Quellen zu kombinieren und gemeinsam zu verwenden” (Statistische Ämter des Bundes und der Länder 2021: 12).

Die verknüpften Daten müssen daher auf einen Linkage-Bias (siehe Abschnitt 3.4) überprüft werden. Dieser darf weder zwischen Subpopulationen noch zwischen geografischen Gebieten vorhanden sein.

Wird anstelle des bundesweiten Bildungsverlaufsregisters das ZSR betrachtet, ist nicht die Qualität des späteren Registers, sondern die Datenqualität der Datengrundlage für das Record-Linkage von

[1] Ferner existieren einige eindeutige Identifikatoren, wie z. B. Telefonnummern oder E-Mail-Adressen, von denen eine Person mehrere besitzen kann. Goldstandard-Daten können anhand dieser Identifikatoren über Record-Linkage erzeugt werden. Die Goldstandard-Daten können dadurch jedoch selektiv sein.

Bedeutung. Üblicherweise wird Datenqualität in Form einer Vielzahl von Dimensionen definiert. Dabei gilt es anzumerken, dass sich keine einheitlichen Definitionen für diese Begriffe in der Literatur finden (Batini & Scannapieco 2006). Darüber hinaus sind nur wenige Dimensionen für das Record-Linkage relevant und können mit den verfügbaren Daten auch untersucht werden. Vier Dimensionen erfüllen diese Anforderungen (Herzog et al. 2007; Wang et al. 2023; Batini & Scannapieco 2006):

- Genauigkeit (Accuracy),
- Vollständigkeit (Completeness),
- Volatilität (Volatility) und
- Relevanz (Relevance).

Unter Genauigkeit soll die Abweichung zwischen der erfassten Ausprägungen eines Merkmals und der wahren Ausprägung verstanden werden (Batini & Scannapieco 2006: 20). Es kann zudem zwischen syntaktischer und semantischer Genauigkeit unterschieden werden (Batini & Scannapieco 2006: 21). Eine Ausprägung ist syntaktisch genau, wenn sie in die Menge der zulässigen Ausprägungen fällt. Ein Geburtsmonat außerhalb des Bereichs 1 bis 12 ist z. B. syntaktisch nicht genau. Semantische Genauigkeit bezeichnet die Nähe zwischen wahrer und erfasster Ausprägung. Ursachen für eine Abweichung können Tippfehler oder eine falsche Erfassung der Ausprägung sein. Die syntaktische Genauigkeit ist damit ein Spezialfall der semantischen Genauigkeit.

Unter Vollständigkeit soll verstanden werden, dass eine Ausprägung oder eine Menge an Ausprägungen tatsächlich erfasst wurde. Diese Dimension ist dadurch mit Nonresponse und Coverage vergleichbar (Herzog et al. 2007: 10). Im Folgenden soll zwischen mehreren Formen der Vollständigkeit unterschieden werden. Die Darstellung ist an Batini & Scannapieco (2006: 23 ff.) angelehnt, wurde jedoch an einigen Stellen ergänzt.

Zunächst sind einzelne Felder vollständig, falls eine Ausprägung erfasst wurde oder ein fehlender Wert die wahre Ausprägung darstellt.[1] Auch innerhalb der Angaben eines Feldes können Teile fehlen. Dies ist z. B. der Fall, wenn bei einer Person mit mehreren Vornamen nicht alle Vornamen erfasst wurden.[2]

Als Maßzahl kann die Vollständigkeit eines ganzen Attributs (QIDs) ermittelt werden. Diese gleicht der Item-Response-Rate und wird als Anteil fehlender Werte eines Attributs m_a an der Gesamtzahl Ausprägungen n berechnet ($c_a = m_a/n$).

Das Gleiche kann auf der Ebene eines Records berechnet werden. Dies wird als Anteil fehlender Werte eines Records m_r an der Gesamtzahl Attribute n_a berechnet ($c_r = m_r/n_a$). Insbesondere die Verteilung dieser Vollständigkeit ist von Bedeutung.

Als letzte Form der Vollständigkeit ist die Coverage zu nennen. Dies ist der Anteil der Zielpopulation, die durch das Register erfasst wurde (Under- und Overcoverage). Die Coverage des ZSR kann mit den zur Verfügung stehenden Daten nicht beurteilt werden. Es kann jedoch davon ausgegangen werden, dass kaum Under- und Overcoverage vorliegen.

[1] In der Literatur wird oft ausgelassen, dass ein fehlender Wert eine gültige Ausprägung darstellen kann. So besitzen z. B. viele Menschen keine zweite Staatsbürgerschaft. In diesem Fall stellt ein fehlender Wert bei der zweiten Staatsbürgerschaft den wahren Wert dar (siehe hierzu auch Schnell 2019b: 46).

[2] Einige Felder können in eine Vielzahl einzelner Felder aufgeteilt werden (z. B. durch die Trennung der Namen). Aus diesem Grund wird das Fehlen einer Angabe innerhalb eines Feldes nicht der Genauigkeit, sondern der Vollständigkeit zugeschrieben. Auf diese Weise bleiben Unterschiede zwischen den Werten immer den gleichen Dimensionen zugeteilt. Würde z. B. ein unvollständiger Name der Dimension der Genauigkeit zugeordnet werden, so würde durch die Trennung der Namensbestandteile ein Problem der Vollständigkeit entstehen. Die Darstellung wäre in diesem Fall inkonsistent.

Volatilität ist eine Dimension, die zumeist als Form der Zeitlichkeit (Timeliness) genannt wird.[1] Bovee et al. (2003) unterscheiden Zeitlichkeit zwischen Alter und Volatilität. Alter beschreibt das Alter der Information, basierend auf dem Zeitpunkt der Erfassung. Volatilität misst die Instabilität eines Merkmals, basierend auf der Häufigkeit von Änderungen. Da das Alter einer Information für die in dieser Arbeit betrachteten Daten nicht zur Verfügung steht, kann nur die Volatilität untersucht werden. Ferner hat das Alter der in dieser Arbeit betrachteten Informationen höchstwahrscheinlich keinen Einfluss auf die Linkage-Qualität.

Relevanz beschreibt die Nützlichkeit und Anwendbarkeit der Daten für eine bestimmte Aufgabe (Wang et al. 2023; Herzog et al. 2007: 8). In Bezug auf Record-Linkage betrifft dies QIDs, welche die Linkage-Qualität wenig oder gar nicht verbessern und daher nicht von Relevanz für das Record-Linkage sind. Nicht relevante QIDs gilt es im Rahmen der Datensparsamkeit vom Linkage-Prozess auszuschließen.

Es finden sich in der Literatur keine Studien über die Datenqualität von QIDs in amtlichen Daten. Eine Annäherung liefert Wurdeman (1993), welcher die Qualität der Texteingabe des US-Zensus 1990 auf Grundlage einer 1 % Stichprobe untersucht. Bei der Codierung des Fragebogens nach Ethnizität stellte er eine durchschnittliche Fehlerquote von 0.54 % fest. Dabei hatten Menschen indianischer Herkunft mit 0.64 % einen signifikant höheren Anteil Fehler als der Rest der Bevölkerung. Zudem konnten signifikante regionale Unterschiede gefunden werden. Adressangaben waren in 0.48 % der Fälle falsch. Ursachen für die Fehler waren unter anderem Eigenheiten bei der Erfassung im Zensus, was einen Vergleich erschwert. Die durchschnittliche Fehlerquote der Codierer betrug 0.19 %.

3.4. Linkage-Bias

Ein Linkage-Bias liegt vor, falls ein Record-Linkage-Verfahren spezifische Teile der Population schlechter verknüpft (Kvalsvig et al. 2019). Bei der Simulation des bundesweiten Bildungsverlaufsregisters konnte gezeigt werden, dass einige Verfahren insbesondere weibliche Migranten schlechter verknüpfen (Weiand 2022). Ursachen für einen Linkage-Bias sind Information-Bias und Verwechslungen (Doidge & Harron 2019; Rothman & Greenland 2014). Ein Information-Bias liegt vor, falls sich die Fehlerarten und -quoten zwischen den Subgruppen unterscheiden (Bohensky 2016). Bei Namen von Migranten kann dies z. B. auftreten, wenn die Namen sich von der mehrheitlichen Bevölkerung unterscheiden und daher weniger geläufig sind (DuVall et al. 2010; Campbell 2009; McCusker et al. 2012).

Verwechslungen liegen vor, falls – gegeben aller Informationen (QIDs) – zwei Records nicht unterschieden werden können oder sogar verwechselt werden (Doidge & Harron 2019). Dies trifft besonders bei Mehrlingen zu (Bentley et al. 2012; Campbell 2009). So können insbesondere Mehrlinge im Kindesalter oft nur anhand des Vornamens unterschieden werden. Auch häufige Namenskombinationen (wie z. B. Peter Müller) können zu Verwechslungen führen, falls nicht genügend Informationen zur Trennung der Fälle vorliegen. Das Risiko für solche Verwechslungen steigt mit der Größe des Datensatzes (Harron et al. 2017; Kvalsvig et al. 2019; Schnell 2022).

Falls die Namensvergabe innerhalb einer Subgruppe homogener als in der restlichen Population ist, kann auch ein Linkage-Bias entstehen. Dies zeigen DuVall et al. (2010) anhand der Verknüpfung von etwa 1.4 Mio. Records einer Genealogie-Datenbank des US-Bundesstaats Utah mit Krankenhausdaten. Bei der Untersuchung von Duplikaten (Mehrfachverknüpfungen) hatten Hispanoamerikaner, obwohl sie nur eine Minderheit im Bundesstaat ausmachen, eine deutlich höhere Wahrscheinlichkeit für Duplikate. Ursache war eine homogenere Namensvergabe innerhalb dieses Bevölkerungsteils.

[1] Viele Definitionen von Timeliness lassen sich besser mit Pünktlichkeit übersetzte, die vorgestellte Definition von Bovee et al. (2003) fällt jedoch nicht darunter, weshalb Zeitlichkeit gewählt wurde.

Eine weitere Ursache für Verwechslungen sind volatile Merkmale. So kann der höhere Anteil Frauen, die bei einer Heirat den Nachnamen ändern, zu unterschiedlichen Verknüpfungsraten und somit einem Linkage-Bias führen (Bohensky et al. 2011; Weiand 2022). Falls Umzüge zu Problemen bei der Verknüpfung führen und bestimmte Subgruppen häufiger umziehen, kann dies auch zu einer Verwechslung und somit zu einem Linkage-Bias führen.

Anhand von Goldstandard-Daten kann der Linkage-Bias durch einen Vergleich der gelinkten und der tatsächlichen Population untersucht werden. Ein univariater Bias kann über eine standardisierte Mittelwertdifferenz (Effektstärken) bestimmt werden (Harron et al. 2017; Cohen 1988). Insbesondere logistische Regressionen können für einen multivariaten Vergleich der Gruppen verwendet werden (Harron et al. 2017; Bopp et al. 2010; Moore et al. 2012; Bohensky et al. 2011; Kvalsvig et al. 2019).[1]

[1] Zum Ausgleich des Linkage-Bias wurden einige Gewichtungsverfahren entwickelt (z. B. Kim & Chambers 2012; Chipperfield & Chambers 2015). Diese Verfahren befinden sich jedoch in den Anfängen. Somit ist ein Verlass auf solche Verfahren beim aktuellen Entwicklungsstand nicht ratsam (Kvalsvig et al. 2019).

4. Datenbeschreibung

Diese Arbeit verwendet Goldstandard-Daten, welche auf Basis eines anderen Record-Linkage-Verfahrens entstanden sind und nachträglich validiert werden (siehe hierzu Abschnitt 4.1). Aus der bisherigen Betrachtung entstehen einige Probleme, welche bei der Interpretation der Ergebnisse im Rahmen eines bundesweiten Bildungsverlaufsregisters beachtet werden müssen:

1. Stichprobe: Hamburg ist keine Zufallsstichprobe der deutschen Bevölkerung. Die Annahmen zur Bildung von Konfidenzintervallen sind somit nicht erfüllt (Berk et al. 1995). Aus diesem Grund werden im Folgenden keine Konfidenzintervalle angegeben und keine Signifikanztests durchgeführt. Nur eine Deskription der Daten ist möglich.
2. Maßstab: Die Simulation des Registers zeigt, dass die Linkage-Qualität von der Größe des Registers abhängt. Je größer das Register, desto schlechter wird die Linkage-Qualität (Schnell 2022). Dies ist auf eine höhere Wahrscheinlichkeit für Records mit gleichen oder sehr ähnlichen QIDs zurückzuführen (Verwechslungen). Da das Hamburger Schülerregister deutlich kleiner ist als ein bundesweites Bildungsverlaufsregister, ist die Verwechslungswahrscheinlichkeit deutlich geringer.
3. Reliabilität: Die Klassifikation der Personen im Datensatz kann Fehler enthalten. Die Linkage-Ergebnisse sowie die Datenqualitätsauswertung kann somit Fehler enthalten.
4. Datenqualität: Die Ergebnisse spiegeln nur die Datenqualität des ZSR wider. Die Datenqualität eines bundesweiten Registers kann sich insbesondere auf Bundeslandebene unterscheiden.
5. QIDs: Nicht alle QIDs standen zur Verfügung, welche im Rahmen eines bundesweiten Registers diskutiert werden.

Im folgenden Abschnitt werden zunächst die Datenquellen und die Aufbereitung der Datensätze beschrieben. Anschließend folgen deskriptive Angaben über die einzelnen Datensätze und ein Vergleich beider Schuljahre. Darauf folgt eine Untersuchung der möglichen Fehlerursachen sowie eine detaillierte Beschreibung der Fehler.

4.1. Beschreibung der Datenquellen

4.1.1. Zentrales Schülerregister (ZSR)[1]

Das ZSR erfasst jedes in Hamburg lebende schulpflichtige Kind, unabhängig davon, ob die Schule des Kindes in Hamburg liegt.[2] Darüber hinaus erfasst das ZSR alle Personen, die in Hamburg eine allgemeinbildende oder berufsbildende Schule besuchen, unabhängig davon, ob eine Person schulpflichtig

[1] Für die hilfreichen Anmerkungen und Erklärungen zu diesem Abschnitt danke ich Herrn Andreas Matzke, Fachliche Leitstelle ZSR.

[2] Nach § 37 HmbSG Absatz 3 dauert die Schulpflicht 11 Schulbesuchsjahre und endet spätestens mit der Vollendung des 18. Lebensjahres.

ist oder in Hamburg lebt (Behörde für Schule und Berufsbildung 2021). Die für diese Arbeit gelieferten Datensätze enthalten nur Personen, die in dem Register als aktive Schüler gekennzeichnet wurden.

Die zentrale Aufgabe des ZSR ist die Feststellung der Erfüllung der Schulpflicht aller in Hamburg lebenden Kinder. Angesichts dessen existieren zwei Datenquellen für das ZSR: die Schulen und das Melderegister. Durch das Hamburger Melderegister werden dem ZSR die Daten aller schulpflichtigen und zukünftig schulpflichtigen Kinder mitgeteilt.[1] Die Schulen übermitteln dem ZSR die Angaben aller an der Schule gemeldeten Schüler (Behörde für Schule und Berufsbildung 2021). Anschließend erfolgt ein Abgleich der Daten. Die Angaben von schulpflichtigen Kindern werden aus dem Melderegister übernommen. Bei allen anderen durch die Schulen gemeldeten Kindern werden die Angaben der Schule übernommen. Für schulpflichtige Kinder, die nicht durch eine Schule gemeldet wurden, wird ein Verfahren zur weiteren Prüfung der Schulpflicht in Auftrag gegeben.

Ein wesentlicher Bestandteil des ZSR ist die GUID (Global Unique Identifier). Diese ermöglicht eine Identifikation einzelner Schüler über einzelne Jahre und Datenquellen hinweg. Die GUID wird in einem halb automatischen Verfahren den Schülern zugeordnet. Grundsätzlich handelt es sich bei dieser Zuordnung um ein Record-Linkage-Problem. Im Gegensatz zu den in dieser Arbeit vorgestellten Record-Linkage-Verfahren beinhaltet das Verfahren zur Erzeugung der GUID einen großen Anteil an manueller Prüfung. So erfolgt eine automatische Zuordnung der GUID nur, wenn bestimmte Merkmalskombinationen (Matchkeys) exakt mit Schülern übereinstimmen, die bereits im Register enthalten sind. Schüler, die nicht auf diese Weise gefunden werden, werden anschließend manuell zugeordnet – sofern diese sich bereits im Register befinden. Dabei gilt es zu beachten, dass die Daten einer Person zwei Jahre im ZSR gespeichert werden, auch wenn eine Person nicht mehr eine Schule in Hamburg besucht. Falls eine Person nach über zwei Jahren Abwesenheit wieder eine Schule in Hamburg besucht, erhält diese eine neue GUID. Dies führt dazu, dass die Anzahl Beobachtungen im ZSR nicht mit der Zeit anwächst.

Aus dem beschriebenen Design resultieren mehrere Schwierigkeiten bei der Datenanalyse. Ein wesentliches Problem besteht darin, dass das ZSR primär als Querschnitt konzipiert ist. Aus diesem Grund ist der Zeitpunkt des Datenabzugs von zentraler Bedeutung. So war ursprünglich ein Vergleich zwischen den Schuljahren 2021/22 und 2022/23 angestrebt. Aufgrund eines verspäteten Abzugs des zweiten Datensatzes muss jedoch das Schuljahr 2023/24 verwendet werden, welches rechtzeitig abgezogen werden konnte.

Die in dieser Arbeit dargestellten Ergebnisse hängen im Wesentlichen von der Gültigkeit der GUID ab. Die GUID wurde nach der FP-Null Regel erstellt. Aufgrund der manuellen Überprüfung und des hohen Anforderungsstandards kann davon ausgegangen werden, dass die GUID keine falsch positiven Zuordnungen verursacht. Da eine Konzentration auf die Vermeidung von falsch positiven Zuordnungen oft mit einer Erhöhung der falsch negativen einhergeht, stellen diese somit das größere Problem dar. Insbesondere jene Fälle, welche aufgrund des Zweijahresabstands technisch bedingt eine andere GUID erhalten haben, sind problematisch. Diese Fälle können jedoch durch ein Clerical-Review weitestgehend aufgelöst werden (siehe Abschnitt 4.2).

Auch mit dem Clerical-Review und dem ursprünglichen Verfahren zur Generierung der GUID können weiterhin Verbindungen zwischen Personen fehlen. Aufgrund des hohen Anspruchs und manuellen Aufwands kann jedoch davon ausgegangen werden, dass die Zahl der falsch negativen Zuordnungen extrem gering ist. Die präsentierten Ergebnisse beinhalten somit wahrscheinlich einen kleinen Fehler, welcher jedoch nicht die Gültigkeit der Ergebnisse beeinträchtigt.

[1] Neben schulpflichtigen Kindern erhält das ZSR auch im Rahmen der 4 1/2-jährigen Testung die Daten für alle zukünftig schulpflichtigen Kinder.

Variable	ZSR	IVDS
GUID	X	X
Vorname	X	
Nachname	X	
Geschlecht	X	X
Geburtstag	X	
Geburtsmonat	X	X
Geburtsjahr	X	X
Straße	X	
Hausnummer	X	
Postleitzahl	X	
1. Staatsangehörigkeit	X	X
2. Staatsangehörigkeit	X	X
Geburtsland		X
Erste Familiensprache		X
RISE-Status		X
Schulform		X
Datenquelle (nur 23/24)	X	

Tabelle 4.1.: Verfügbare Variablen im ZSR und IVDS.

Die Variablen, die aus dem ZSR zur Verfügung stehen, werden in Tabelle 4.1 aufgelistet. Eine Angabe über die Datenquelle (Schule oder Melderegister) konnte nur für das Schuljahr 2023/24 geliefert werden.

4.1.2. Schülerindividualdatensatz (IVDS)

Der IVDS – auch bekannt als Schuljahreserhebung – ist eine jährliche Erhebung der Behörde für Schule und Berufsbildung Hamburg (BSB), die nach Schuljahresbeginn an allen Hamburger Schulen erhoben wird. Stichtag für die allgemeinbildenden Schulen war der 13.09.2021 bzw. der 22.09.2023 und für berufsbildende Schulen der 20.10.2021 bzw. der 05.10.2023 (Behörde für Schule und Berufsbildung 2022; Behörde für Schule und Berufsbildung 2024). Dabei werden eine Vielzahl von Merkmalen erhoben, die besonders zu Prognosezwecken, interner Bedarfsplanung und zur Beantwortung parlamentarischer Anfragen genutzt werden.[1] Der IVDS ist zudem die Grundlage für die Berichtspflicht an die KMK und das Statistische Bundesamt zur bundesweiten Schulstatistik (Behörde für Schule und Berufsbildung 2022; Statistisches Bundesamt 2023).

Tabelle 4.1 listet alle Variablen auf, die aus dem IVDS geliefert wurden. Das Geburtsdatum (Monat und Jahr), die Staatsangehörigkeit und das Geschlecht sind sowohl für das ZSR als auch für den IVDS verfügbar. Es gilt hervorzuheben, dass beide Datensätze auf den Angaben der Schulen basieren. Sofern im ZSR die Daten über eine Person von der Schule stammen, lassen sich daher nur sehr wenige Unterschiede zwischen den Merkmalen im IVDS und ZSR feststellen. Der Zusammenhang dieser beiden Datensätze zeigt sich auch darin, dass die GUID Bestandteil des IVDS ist. Die GUID stammt dabei aus dem ZSR und wird im digitalen Verwaltungssystem der Hamburger Schulen (DiViS) mitgeführt und für den IVDS exportiert. Somit können dem ZSR weitere Informationen aus dem IVDS angefügt werden.

Aus dem IVDS wurden alle aktiven Schüler sowie die Schulentlassenen der allgemeinbildenden Schulen geliefert. Da das ZSR nur aktive Schüler beinhaltet, befinden sich jeweils etwa 80 % der Fälle des IVDS

[1] https://www.hamburg.de/schulstatistiken/ (abgerufen am 12.12.2023).

im ZSR. Da das ZSR auch Schüler in berufsbildenden Schulen enthält, finden sich jeweils etwa 90 % der Personen des ZSR im IVDS.

Die gelieferten Variablen enthalten auch den RISE-Status. RISE (Rahmenprogramm Integrierte Stadtteilentwicklung) ist ein Index zur Einstufung der Lebensqualität in Hamburger Wohnquartieren (Behörde für Stadtentwicklung und Wohnen 2022). Der Wertebereich ist ganzzahlig und liegt zwischen 1 (sehr niedrig) und 4 (hoch). Der Index wird aus einer Vielzahl an Informationen berechnet, die an dieser Stelle nicht weiter aufgeführt werden sollen. Es gilt anzumerken, dass der Index nicht gleich verteilt ist. So wurden 2021 66.1 % der statistischen Gebiete dem Status 3 (Mittel) zugeordnet. Weitere 16.8 % dem Status 4 und nur 8.3 % bzw. 8.8 % dem Status 1 und 2 (Behörde für Stadtentwicklung und Wohnen 2021: 12). Damit lebten 2021 81.4 % der Hamburger Bevölkerung in statistischen Gebieten mit einem RISE-Status von 3 oder 4 (Behörde für Stadtentwicklung und Wohnen 2021: 13). Im Rahmen dieser Arbeit eignet sich der Index als Indikator für den sozialen Status einer Person.

4.2. Datenaufbereitung

Für einen optimalen Verknüpfungserfolg müssen alle QIDs in eine möglichst gleiche Form übertragen werden (Herzog et al. 2007; Bohensky 2016). Dabei ist eine Bereinigung von Sonderzeichen und Diakritika (z. B. Akzente) hilfreich, da diese oft Fehler beinhalten, jedoch nur einen geringen Informationsgehalt haben (Kaalep et al. 2022). Eine Umwandlung z. B. des Namens RENÈ in RENE erhöht somit die Datenqualität, da der Akzent häufig Fehler aufweist, weil er z. B. vergessen oder in die andere Richtung geschrieben wird. Zur Bereinigung wurden daher alle Namen und Straßen in Großbuchstaben umgewandelt und alle Buchstaben in das lateinische Alphabet übertragen (z. B. À zu A und Ö zu OE). Bindestriche wurden in Leerzeichen umgewandelt. Alle weiteren Sonderzeichen wurden entfernt. Namen und Straßen bestehen somit nur noch aus den Buchstaben A-Z und Leerzeichen. Ferner wurden die Staatsangehörigkeiten, Geburtsländer und Familiensprache einheitlich codiert.

Allen Schülern wurden die Identifikatoren verschiedener geografischer Rastergitter nach dem INSPIRE-Referenzsystem angehängt (Gitter-ID). Rastergitterdaten für Deutschland werden durch das Bundesamt für Kartografie und Geodäsie (BKG) zur Verfügung gestellt.[1] Verwendet wurden die Auflösungen 100 m, 250 m, 500 m und 1 km in einer Lambertschen flächentreuen Azimutalprojektion (LAEA, Lambert azimuthal equal-area projection) im ETRS89 Datum. Zur Feststellung einer Gitter-ID wurden zunächst die geografische Position des Wohnorts bestimmt.[2] Dies erfolgte durch eine Verknüpfung der Adressdaten mit aktuellen Adressdaten aus OpenStreetMap (OSM).[3] Hierzu wurden alle bekannten Hausnummern in Deutschland mit zugehörigem Straßennamen und Postleitzahl extrahiert.[4] Es wurden alle Adressen in Deutschland extrahiert, da einige im ZSR geführte Schüler nicht im Hamburg wohnen (siehe Abschnitt 4.1). Es gilt anzumerken, dass die OSM-Daten nicht vollständig sind und auch Datenfehler enthalten können.

Die OSM-Daten wurden zunächst exakt mit den Adressdaten der Schüler (Straße, Hausnummer mit Zusatz und Postleitzahl) verknüpft.[5] Anschließend erfolgte eine weitere Verknüpfung anhand der Haus-

[1] https://gdz.bkg.bund.de/index.php/default/inspire/sonstige-inspire-themen/geographische-gitter-fur-deutschland-in-lambert-projektion-geogitter-inspire.html (abgerufen am 09.02.2024)

[2] Die Gitter-ID wird aus den Koordinaten der unteren linken Ecke der Gitterzelle gebildet. Daher wurde aus den LAEA-Koordinaten die entsprechende Gitter-ID durch Abrunden der x- und y-Koordinaten gebildet.

[3] https://download.geofabrik.de/europe/germany.html (Datenstand 01.02.2024)

[4] OpenStreetMap führt die Attribute Straße und Platz. Beide Attribute wurden verwendet, wobei der Name des Platzes verwendet wurde, falls der Straßenname fehlte.

[5] Da häufig eine Adresse nicht auffindbar war, weil die Hausnummer in den OSM-Daten fehlte, wurde auf ein probabilistisches Verfahren bei der Verknüpfung der Daten verzichtet.

nummer ohne Zusatz. Alle Schüler, deren Koordinaten nicht exakt ermittelt werden konnten, mussten ungenau verknüpft werden. Zunächst erfolgte die Verknüpfung bei exakter Übereinstimmung der Postleitzahl und ungenauer Übereinstimmung des Straßennamens (Jaro-Winkler-Ähnlichkeit mit Schwellenwert 0.8). Dies bedeutet, dass Straßen mit Tippfehlern oder anderer Schreibweise ermittelt wurden.[1] Es wurde immer die Straße mit der höchsten Übereinstimmung verwendet. Anhand der gewählten Straße wurden die Koordinaten der Hausnummer ermittelt. Hierbei wurde das nächstgelegene bekannte Gebäude zur gegebenen Hausnummer gesucht (absolute nummerische Distanz). Dies ist notwendig, da die Hausnummer für einige Gebäude in den OSM-Daten fehlen. Alle Schüler, deren Adressen weiterhin nicht ermittelt werden konnten, durchliefen ein ähnliches Verfahren, bei dem jedoch nur die ersten zwei Stellen der Postleitzahl übereinstimmen mussten. Dies betraf besonders Adressen an Grenzen von Postleitzahl-Gebieten, da die OSM-Daten hier eine leichte Ungenauigkeit aufwiesen.[2] Bei jeder Verknüpfungsmethode erhielt die Person die Geo-Koordinate der ermittelten Hausnummer.[3]

Die Verknüpfungsraten für beide Datensätze waren fast identisch. Für das Schuljahr 2023/24 konnten 87.8 % der Adressen direkt ermittelt werden, 4.1 % durch Auslassen des Zusatzes, 7.4 % durch ungenauen Straßennamen und nahe gelegener Hausnummer sowie weitere 0.6 % durch eine ungenaue Postleitzahl. Insgesamt 892 Adressen konnten nicht identifiziert werden. Die Koordinaten für 745 dieser Adresse konnte über Google-Maps manuell ermittelt werden und wurde den Daten hinzugefügt.[4] Anhand der ermittelten Koordinaten wurde anschließend der Name des entsprechenden Hamburger Stadtteils den Personen zugeordnet[5] sowie eine Postleitzahl, welche über die Zuordnung der Koordinaten und der Postleitzahlgebiete gewonnen wurde (Datenquelle OSM).[6]

In der Simulation wurde angenommen, dass Migranten eine doppelt so hohe Fehlerquote besitzen als Einheimische (Schnell 2022). Ferner zeigen die in Kapitel 2 vorgestellten Erwartungen an ein bundesweites Bildungsverlaufsregister, dass es möglich sein sollte, Bildungsverläufe nach Herkunft zu vergleichen. Zur Evaluation dieser Punkte wurde jeder Person einer Herkunft zugeordnet. Unterschieden wird zwischen Deutschen, Migranten und Menschen mit Migrationshintergrund.

Migranten werden als Personen definiert, die nicht in Deutschland geboren wurden und keine deutsche Staatsbürgerschaft besitzen. Als Menschen mit Migrationshintergrund werden Personen definiert, die neben einer ausländischen Staatsbürgerschaft eine deutsche Staatsbürgerschaft besitzen oder die in Deutschland geboren wurden und mindestens eine andere Staatsbürgerschaft besitzen. Als Deutsche werden Personen definiert, die ausschließlich eine deutsche Staatsbürgerschaft besitzen und in

[1] Ein besonders häufiger Fehler bestand darin, dass eine Straße, die mit “Straße” endete, als “Str.” abgekürzt und ein Leerzeichen davor manchmal fehlte bzw. eingesetzt wurde. Bei der Datenaufbereitung wurden die Straßennamen in ein einheitliches Format (Abkürzung ohne führendes Leerzeichen) übertragen.

[2] Straßen mit den Namen UNBEKANNT, WURDE BEI ANMELDUNG, MUSTER, KOMMT IN KUERZE, POSTFACH, PO BOX und XX wurden durch einen fehlenden Wert ersetzt.

[3] Die zugeordnete Adresse ist eine Punkt-Koordinate. Einige Hausnummern in den OSM-Daten werden als Punkt-Koordinate und andere als Polygon (Grundfläche des Gebäudes) geführt. Falls ein Polygon vorlag, wurde der geometrische Schwerpunkt (Centroid) als Punkt-Koordinate der Adresse verwendet.

[4] Die Ermittlung der Adresse über Google-Maps erfolgte anhand eines Skripts, welches automatisch die bekannten Adressangaben zusammenfügte und in einem privaten Browserfenster bei Google-Maps aufrief. Konnte nicht sofort eine Adresse gefunden werden, erfolgten weitere Abfragen durch Auslassen der Postleitzahl und durch Angabe des Ortsnamens. Auffällig waren insbesondere Adressen, die auf Kleingartenvereine referenzierten.

[5] https://geodienste.hamburg.de/download?url=https://geodienste.hamburg.de/HH_WFS_Verwaltungsgrenzen&f=json (abgerufen am 05.01.2022)

[6] https://public.opendatasoft.com/api/explore/v2.1/catalog/datasets/georef-germany-postleitzahl/exports/geojson?lang=en&timezone=Europe%2FBerlin(abgerufen am 05.01.2022)

Deutschland geboren wurden. Die Herkunft wurde anhand aller vorhandenen Merkmale aus dem ZSR und IVDS (Staatsangehörigkeit und Geburtsland) gebildet.[1]

Bei der erstmaligen Verknüpfung des Datensatzes mit den getesteten Record-Linkage-Verfahren (Abschnitt 5) wurden einige auffällige, falsch positive Verknüpfungen entdeckt. Hierzu zählten unter anderem Personen, welche sich in allen Merkmalen glichen, jedoch unterschiedliche GUIDs besaßen. Diese Fälle sind – wie in Abschnitt 4.1 beschrieben – aufgrund des Designs des ZSR möglich. Angesichts dessen musste ein Clerical-Review durchgeführt werden. Es wurden nur Szenarien verwendet, bei denen das volle Geburtsdatum und der vollständige Name vorlag (Szenario 1-10 Abschnitt 5.2). Bei Szenarien, in denen eine volle Adresse oder Gitter-ID vorlag, wurden eindeutige exakte Verknüpfungen mit unterschiedlichen GUIDs als true positiv neu klassifiziert. Darüber hinaus wurde die Anzahl Szenarien gezählt, bei denen ein Paar probabilistisch oder exakt verknüpft wurde. Paare, die in über 10 Fällen verknüpft wurden, bei denen Vorname und Geburtsjahr übereinstimmte sowie Paare, die in über drei Fällen verknüpft wurden und Straße und Vorname exakt übereinstimmten, wurden auch als true positiv neu klassifiziert. Bei Betrachtung der Daten zeigten sich keine auffälligen Unterschiede zwischen allen auf diese Art identifizierten Record-Paaren. Es wurde somit angenommen, dass echte false positives in diesen Fällen höchst unwahrscheinlich sind. Insgesamt 412 Personen wurden direkt identifiziert und neu klassifiziert.

Paare, die in mehr als 4 Szenarien verknüpft wurden, wurden manuell durch zwei Personen getrennt codiert. Beide erhielten die Paare in einer anderen, zufälligen Reihenfolge. Die Intercoder-Reliabilität betrug $\kappa = 0.8$. Unterschiedliche Codierungen wurden unter gemeinsamer Absprache endgültig klassifiziert. Durch dieses Verfahren wurden weitere 194 Paare neu als true positive klassifiziert. Die Klassifikation erfolgte möglichst konservativ.

Alle neu klassifizierten Paare erhielten neue, klar unterscheidbare GUIDs. Insgesamt können durch diese Datenaufbereitung $182,385$ Personen über die GUID verknüpft werden. Davon sind $21,226$ (11.6 %) Migranten, $37,341$ (20.5 %) Menschen mit Migrationshintergrund und $123,818$ (67.9 %) Deutsche.

Da die Datenquellen für das ZSR (Schule und Melderegister) nur für das Schuljahr 2023/24 zur Verfügung stand, wurden die Werte für das Schuljahr 2021/22 deduktiv imputiert (Lohr 2021: 335). Dies ist möglich, da die Datenquelle relativ leicht regelbasiert abgeleitet werden können: Alle Daten stammen von den Schulen, mit Ausnahme von schulpflichtigen in Hamburg lebenden Kindern. Die Schulpflicht wurde über das Alter (4-18 Jahre) festgestellt und der Wohnort über die Postleitzahl. Ferner wies der Datensatz für das Schuljahr 2023/24 30 fehlende Angaben zur Datenquelle auf. Diese wurden nach dem gleichen Verfahren imputiert.

Abbildung 4.1 zeigt den Anteil der durch die Schule gemeldeten Fälle nach Geburtsmonat für beide Datensätze. Die Ausreißer lassen sich durch eine sehr geringe Zahl Beobachtungen in dem Zeitraum begründen und sind somit auf den Zufall zurückzuführen. Auffällig sind zudem Personen, die jünger als 4 Jahre sind. Personen in diesem Alter sollten nicht durch das ZSR erfasst werden, sodass diese Beobachtungen höchstwahrscheinlich auf Datenfehler zurückzuführen sind. Die Abbildung zeigt, dass die Imputation der Datenquelle für das Schuljahr 2021/22 erfolgreich war. So ist der Anteil unter

[1] Eine Klassifikation der Herkunft könnte zusätzlich auch die Familiensprache berücksichtigen. Hierzu wurde die zusätzliche Bedingung getestet, dass die Familiensprache Deutsch zur Klassifikation von Deutschen notwendig ist. In Deutschland geborene Menschen mit deutscher Staatsbürgerschaft, die jedoch nicht Deutsch als Familiensprache angegeben haben, wurden den Menschen mit Migrationshintergrund zugerechnet. Eine solche Klassifikation erhöhte den Anteil Menschen mit Migrationshintergrund insgesamt um 5.3 %. Die in dieser Arbeit vorgestellten Ergebnisse veränderten sich jedoch nur geringfügig. Dabei blieben alle Interpretationen erhalten. Daher wurde die sparsamere Definition verwendet. Auf die Verwendung des weichen Merkmals Familiensprache wurde verzichtet und ausschließlich die harten Merkmale Geburtsort und Staatsangehörigkeit verwendet.

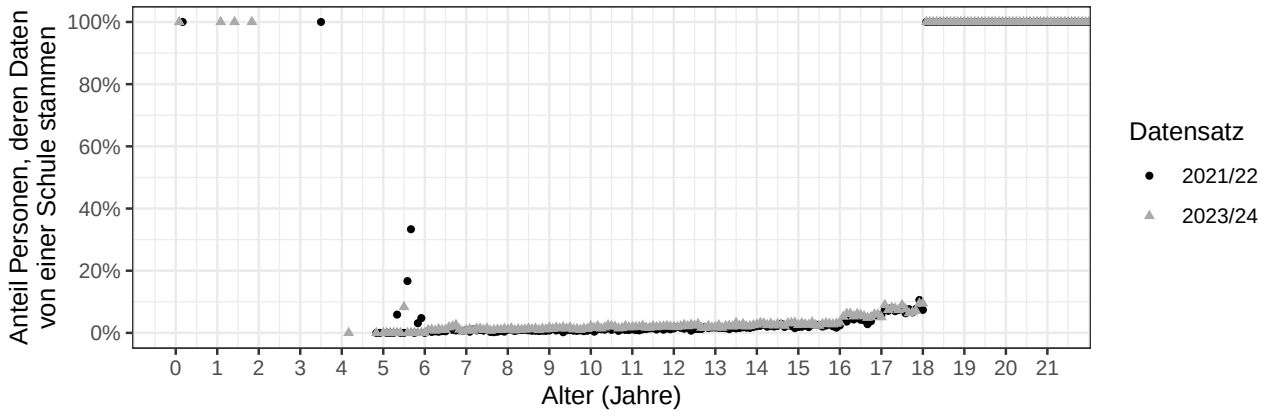


Abbildung 4.1.: Anteil Personen nach Alter (gruppiert nach Monate), deren Daten von einer Schule stammen. Die Abbildung ist rechts zensiert. Ältere Personen werden in beiden Datensätzen alle durch die Schule gemeldet.

18-jähriger Personen, die durch die Schule gemeldet wurden, im imputierten Datensatz nur um 0.6 % geringer als im Datensatz mit den empirischen Werten (Schuljahr 2023/24).

Der RISE-Status wurde für alle berufsbildenden Schüler imputiert, da die entsprechende Information aus dem IVDS für diese Personen nicht vorlag. Hierzu wurde der RISE-Status der Schüler aus dem IVDS verwendet. Da der RISE-Status eine geografische Information ist, wurde zur Imputation die Gitterzellen verwendet. Hierzu wurde mit der kleinsten Gitterzellengröße (100 m) begonnen und danach mit immer größer werdenden Gitterzellengrößen nach einem bekannten RISE-Status gesucht. Die Suche erfolgte dabei getrennt nach Schuljahr. Konnten mehrere Werte für eine Zelle ermittelt werden, wurde der Modalwert gewählt. Die Prozedur wurde mit den Daten des jeweils anderen Schuljahres wiederholt, sodass für möglichst jede Person aus Hamburg der RISE-Status identifiziert werden konnte. Etwa 95 % der Beobachtungen beider Schuljahre konnten mit den Angaben der 100m-Gitterzelle des gleichen Schuljahres imputiert werden. Weitere 4 % konnten bereits mit der 250m-Gitterzelle imputiert werden. Der RISE-Status für 36, 290 Records des Schuljahrs 2021/22 und 33, 193 Records des Schuljahrs 2023/24 wurde insgesamt imputiert.

4.3. Deskriptive Angaben

Zum Verständnis der Datenqualitätseigenschaften werden im Folgenden zunächst die beiden Datensätze getrennt dargestellt. Da diese Arbeit im Wesentlichen auf eine Beschreibung des ZSR abzielt, wird der IVDS nicht getrennt beschrieben. Bei einer Betrachtung von Angaben aus dem IVDS werden Schüler von berufsbildenden Schulen ausgeschlossen. Insgesamt enthält das ZSR für das Schuljahr 2021/22 246, 472 Records und das Schuljahr 2023/24 252, 230 Records. Für das Schuljahr 2021/22 besitzen 199, 426 Records Angaben aus dem IVDS und für das Schuljahr 2023/24 211, 381 Records. Insgesamt finden sich 182, 385 Records in beiden Datensätzen. Davon besitzen 160, 088 Records Angaben aus dem IVDS für beide Schuljahre.

Da die getrennten Datensätze für das spätere Record-Linkage verwendet werden, sollen auch einige für das Record-Linkage relevante Deskriptionen erfolgen. Hierzu zählt zunächst die Anzahl der einzigartigen Werte und somit die Anzahl Ausprägungen, die ein QID annimmt. Je mehr Ausprägungen, desto besser ist das Merkmal für die Verknüpfung geeignet. Neben der Anzahl einzigartiger Werte ist auch die Verteilung dieser Werte von zentraler Bedeutung. Je homogener eine Variable verteilt

ist, desto besser eignet sich das Merkmal für die Verknüpfung. Diese beiden Eigenschaften werden in der Entropie zusammengefasst. Die Entropie H gibt den Informationsgehalt einer Zufallsvariablen X mit den möglichen Ausprägungen x_i und der Wahrscheinlichkeit für die Ausprägung $p(x_i)$ an und ist definiert als

$$H(X) = -\sum_{i=1}^{n} p(x_i) \cdot \log_2 p(x_i) \tag{4.1}$$

(Shannon 1948). Je höher die Entropie, desto höher ist der Informationsgehalt der Variable. Das Maximum der Entropie liegt bei einer Gleichverteilung der Werte vor. In diesem Fall beträgt $H = 1/\log_2 n$. Das Verhältnis der Entropie zur maximalen Entropie wird als normalisierte Entropie bezeichnet und ist definiert als (Kumar et al. 1986):

$$\bar{H}(X) = -\frac{1}{\log_2 n} \sum_{i=1}^{n} p(x_i) \cdot \log_2 p(x_i). \tag{4.2}$$

Die normalisierte Entropie wird im weiteren Verlauf dieser Arbeit verwendet und soll hier zunächst nur definiert werden.

Die Entropie und Anzahl einzigartiger Werte aller Variablen für beide Datensätze werden in Tabelle 4.2 dargestellt. Zusätzlich werden die Anteile und Werte der zwei häufigsten Ausprägungen dargestellt. Die Ergebnisse entsprechen den Erwartungen. So besitzen vor allem Namen und Adressen einen hohen Informationsgehalt. Die 100m-Gitter haben die höchste Entropie aller Adressmerkmale.

Bei einer genaueren Untersuchung der Gitter-IDs zeigt sich, dass im Schuljahr 2021/22 4.7 % eine eindeutige 100m-Gitter-ID besitzen. Der Median der Anzahl Personen mit derselben 100m-Gitter-ID beträgt 14. Bei einem 250m-Gitter besitzen 2.3 % eine eindeutige Gitter-ID (Median 59), bei einem 500m-Gitter 1.4 % (Median 190) und bei einem 1km-Gitter 0.9 % (Median 649).

Sowohl die Staatsbürgerschaft als auch das Geburtsland besitzen trotz vieler Ausprägungen einen geringen Informationsgehalt. Eine Möglichkeit, den Informationsgehalt der Staatsbürgerschaft zu erhöhen, besteht in der Zusammenfassung der Merkmale. Hierzu erhalten alle Personen zunächst den Wert ihrer ersten Staatsbürgerschaft. Ist diese deutsch und besitzen die Personen eine zweite Staatsbürgerschaft, so erhalten sie stattdessen den Wert ihrer zweiten Staatsbürgerschaft. Diese neue Variable wird in Tabelle 4.2 als “Staatsbürgerschaft zu.” aufgeführt. Die Zusammenfassung der Merkmale bringt einen höheren Informationsgehalt mit gleichzeitig wenigen fehlenden Werten.

Gut ersichtlich ist zudem, dass die häufigste Ausprägung des Geburtsmonats und Geburtstags die 1 ist. Dies lässt sich damit begründen, dass typischerweise bei amtlichen Daten der 1.1. als Geburtsmonat eingetragen wird, falls kein exaktes Geburtsdatum vorliegt. Dies ist häufig bei Migranten der Fall.

Weiter auffällig sind die häufigsten Straßennamen. Während die Kielerstr. eine ca. 6 km lange Straße im Zentrum Hamburgs ist, handelt es sich bei dem Achtern Born und dem Fritz-Flinte-Ring jeweils um kurze Straßen in den Großwohnsiedlungen Hamburg-Steilshoop und Osdorfer Born. Es ist zu erwarten, dass Großgebäude und Großwohnsiedlungen aufgrund einer höheren Verwechslungswahrscheinlichkeit mehr Probleme bei der Verknüpfung verursachen.

In Tabelle 4.2 ist zudem erkennbar, dass der Anteil Deutscher zwischen den Schuljahren um etwa 3.4 % abgenommen hat. Ursache für diese Veränderung sind insbesondere ukrainische Kriegsflüchtlinge. So steig der Anteil Personen mit ukrainischer Staatsbürgerschaft (1. Staatsbürgerschaft) von 0.2 % auf 2.3 % zwischen den Schuljahren an.

Tabelle 4.2.: Angaben über die Verteilung der Werte aller Variablen für beide Schuljahre. Fehlende Werte wurden bei der Liste der häufigsten Werte und Anzahl einzigartiger Werte nicht berücksichtigt (n_{unqiue} = Anzahl einzigartiger Ausprägungen; zu. = zusammengefasst; gen. = Variable generiert (siehe Abschnitt 4.2)).

					häufigster Wert		zweithäufigster Wert	
Variable	Quelle	Datensatz	n_{unique}	$H(x)$	Wert	Anteil (%)	Wert	Anteil (%)
Geschlecht	IVDS	2021/22	2	1.00	Männlich	51.11	Weiblich	48.89
Geschlecht	IVDS	2023/24	2	1.00	Männlich	50.94	Weiblich	49.06
Geburtsjahr	IVDS	2021/22	46	3.78	2014	8.92	2013	8.59
Geburtsjahr	IVDS	2023/24	46	3.78	2016	9.13	2015	8.81
Geburtsmonat	IVDS	2021/22	12	3.58	1	9.14	8	8.82
Geburtsmonat	IVDS	2023/24	12	3.58	1	9.07	8	8.88
Geburtsland	IVDS	2021/22	161	1.28	Deutschland	86.42	Syrien	2.04
Geburtsland	IVDS	2023/24	165	1.47	Deutschland	83.74	Ukraine	2.39
1. Staatsbürgerschaft	IVDS	2021/22	145	1.32	Deutschland	85.46	Afghanistan	2.21
1. Staatsbürgerschaft	IVDS	2023/24	148	1.56	Deutschland	81.95	Afghanistan	2.80
2. Staatsbürgerschaft	IVDS	2021/22	155	1.32	Türkei	3.43	Afghanistan	1.00
2. Staatsbürgerschaft	IVDS	2023/24	156	1.30	Türkei	3.11	Afghanistan	0.96
Staatsbürgerschaft zu.	IVDS	2021/22	161	2.36	Deutschland	71.52	Türkei	3.87
Staatsbürgerschaft zu.	IVDS	2023/24	165	2.55	Deutschland	68.45	Afghanistan	3.75
Familiensprache	IVDS	2021/22	117	2.17	Deutsch	71.56	Türkisch	4.80
Familiensprache	IVDS	2023/24	114	2.39	Deutsch	67.89	Türkisch	4.84
RISE-Status	IVDS	2021/22	4	1.68	3	60.13	4	16.48
RISE-Status	IVDS	2023/24	4	1.70	3	58.97	4	17.61
Schultyp	IVDS	2021/22	12	1.76	Grundschule	34.41	Stadtteilschule	33.47
Schultyp	IVDS	2023/24	11	1.76	Grundschule	35.38	Stadtteilschule	32.85
Geschlecht	ZSR	2021/22	2	1.00	Männlich	52.01	Weiblich	47.99
Geschlecht	ZSR	2023/24	2	1.00	Männlich	51.53	Weiblich	48.46
Vorname	ZSR	2021/22	112105	14.81	LEON	0.31	ALEXANDER	0.27
Vorname	ZSR	2023/24	114587	14.88	LEON	0.26	JONAS	0.24
Nachname	ZSR	2021/22	76886	14.89	SCHMIDT	0.35	MUELLER	0.30
Nachname	ZSR	2023/24	80727	14.99	SCHMIDT	0.32	MUELLER	0.27
Geburtstag	ZSR	2021/22	31	4.95	1	4.27	20	3.45
Geburtstag	ZSR	2023/24	31	4.95	1	4.28	20	3.44
Geburtsmonat	ZSR	2021/22	12	3.58	1	9.27	7	8.78
Geburtsmonat	ZSR	2023/24	12	3.58	1	9.21	8	8.85
Geburtsjahr	ZSR	2021/22	63	4.34	2014	7.29	2013	7.04
Geburtsjahr	ZSR	2023/24	60	4.27	2016	7.73	2015	7.49

Tabelle 4.2.: Fortsetzung Angaben über die Verteilung der Werte aller Variablen.

					häufigster Wert		zweithäufigster Wert	
Variable	Quelle	Datensatz	n_{unique}	$H(x)$	Wert	Anteil (%)	Wert	Anteil (%)
1. Staatsbürgerschaft	ZSR	2021/22	162	1.37	Deutschland	85.10	Afghanistan	2.32
1. Staatsbürgerschaft	ZSR	2023/24	165	1.60	Deutschland	81.67	Afghanistan	2.91
2. Staatsbürgerschaft	ZSR	2021/22	161	1.36	Türkei	3.27	Afghanistan	1.02
2. Staatsbürgerschaft	ZSR	2023/24	161	1.38	Türkei	2.98	Afghanistan	1.00
Staatsbürgerschaft zu.	ZSR	2021/22	175	2.42	Deutschland	70.95	Türkei	3.78
Staatsbürgerschaft zu.	ZSR	2023/24	177	2.61	Deutschland	67.84	Afghanistan	3.89
Straße	ZSR	2021/22	14137	12.12	ACHTERN BORN	0.24	KIELERSTR	0.24
Straße	ZSR	2023/24	12277	12.04	KIELERSTR	0.25	FRITZFLINTERING	0.24
Hausnummer	ZSR	2021/22	3268	8.28	1	2.18	3	2.13
Hausnummer	ZSR	2023/24	3275	8.29	1	2.13	3	2.09
Postleitzahl	ZSR	2021/22	1426	6.86	21035	2.21	22159	2.01
Postleitzahl	ZSR	2023/24	1232	6.78	21035	2.17	22159	2.13
Datenquelle ZSR	gen.	2021/22	2	0.75	Melderegister	78.48	Schule	21.52
Datenquelle ZSR	ZSR	2023/24	2	0.71	Melderegister	80.65	Schule	19.35
100m-Gitter-ID	OSM	2021/22	36310	14.22	100mN33867E43275	0.07	100mN33754E43330	0.07
100m-Gitter-ID	OSM	2023/24	34163	14.16	100mN33887E43220	0.10	100mN33867E43275	0.09
250m-Gitter-ID	OSM	2021/22	14402	12.29	250mN338925E432475	0.17	250mN337500E432225	0.15
250m-Gitter-ID	OSM	2023/24	12960	12.23	250mN338925E432475	0.16	250mN338325E431450	0.15
500m-Gitter-ID	OSM	2021/22	7678	10.70	500mN33890E43245	0.41	500mN33750E43220	0.38
500m-Gitter-ID	OSM	2023/24	6739	10.62	500mN33890E43245	0.41	500mN33750E43220	0.36
1km-Gitter-ID	OSM	2021/22	4458	9.03	1kmN3389E4324	0.79	1kmN3385E4318	0.73
1km-Gitter-ID	OSM	2023/24	3813	8.95	1kmN3389E4324	0.78	1kmN3385E4318	0.73

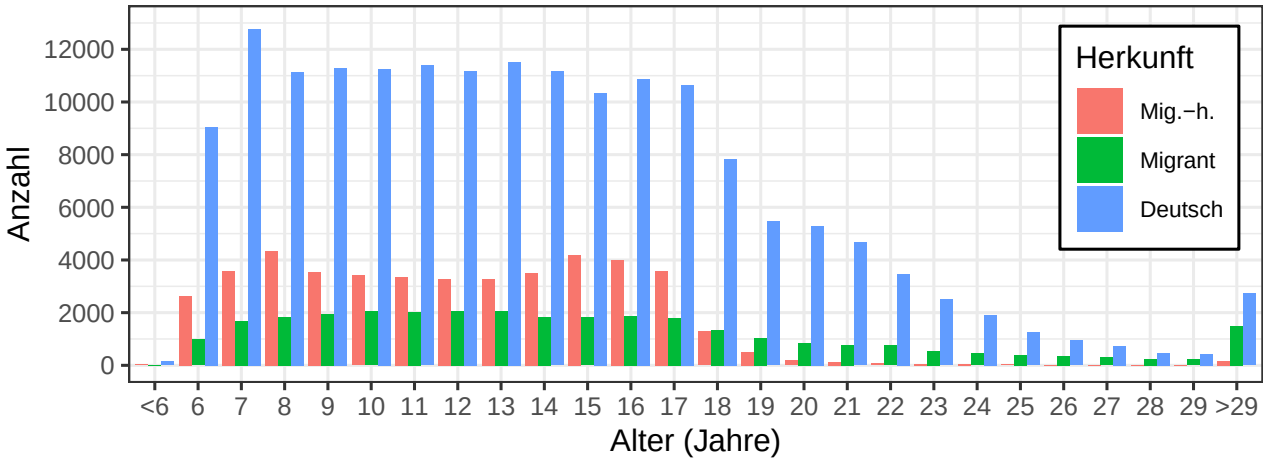


Abbildung 4.2.: Anzahl Beobachtungen nach Alter und Herkunft für das Schuljahr 2021/22 (Mig.-h. = Migrationshintergrund).

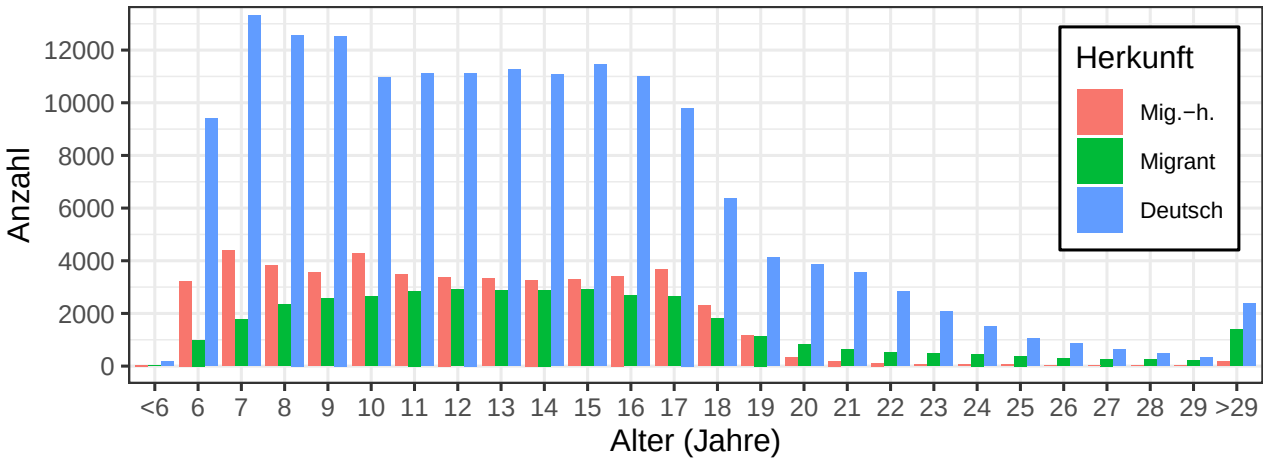


Abbildung 4.3.: Anzahl Beobachtungen nach Alter und Herkunft für das Schuljahr 2023/24 (Mig.-h. = Migrationshintergrund).

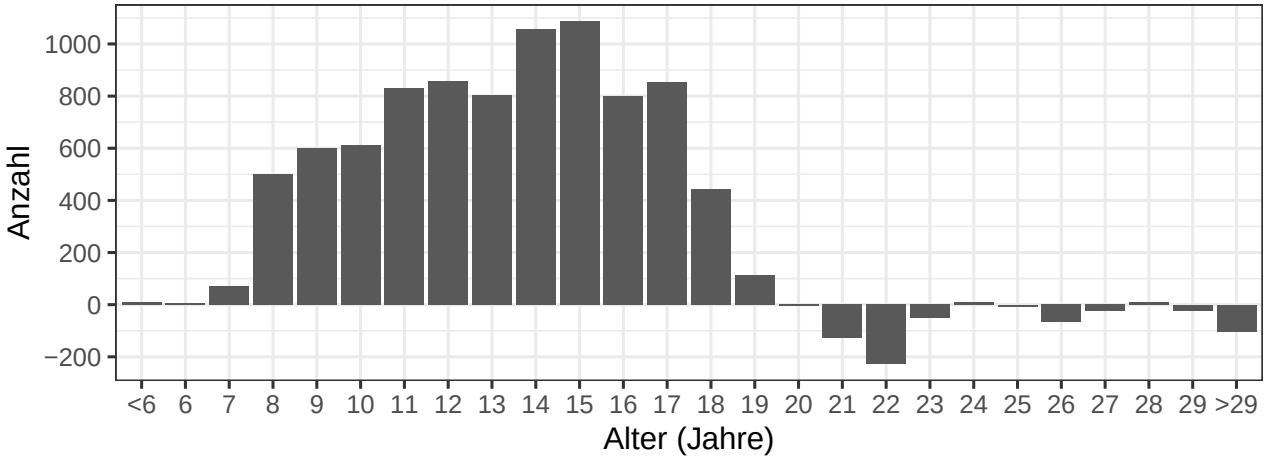


Abbildung 4.4.: Veränderung der Gesamtzahl Migranten für das Schuljahr 2023/24 im Verhältnis zum Schuljahr 2021/22 nach Alter.

Diese Veränderung lässt sich auch in der Altersverteilung der Records erkennen. Abbildungen 4.2 und 4.3 stellen die Anzahl Records nach Alter und Herkunft dar. Die Abbildung 4.3 zeigt einen deutlichen Anstieg der Anzahl Migranten für alle Altersgruppen. Die Veränderung wird in Abbildung 4.4 dargestellt. Abbildung 4.4 zeigt einen klaren Zuwachs von Migranten im schulpflichtigen Alter, besonders ab einem Alter von 8 Jahren. Da ein Großteil des Zuwachses an Migranten auf ukrainische Kriegsflüchtlinge zurückzuführen sind, muss beachtet werden, dass Männer ab 18 Jahren die Ukraine nicht verlassen durften. Dies erklärt zum Teil, dass kein Zuwachs bei älteren Migranten zu beobachten ist.

Die Abbildungen 4.2 und 4.3 zeigen ferner, dass die meisten Beobachtungen von Kindern im schulpflichtigen Alter stammen. So nimmt die Anzahl Beobachtungen ab einem Alter von 18 Jahren stark ab. Auffällig ist dabei, dass die Anzahl Menschen mit Migrationshintergrund deutlich schneller abnimmt als die Anzahl Deutscher und Migranten. Für das Schuljahr 2021/22 beträgt beispielsweise der Anteil Menschen mit Migrationshintergrund 22.3 % bei 17-Jährigen und nur 2.9 % bei 20-Jährigen. Diese Bevölkerungsgruppe ist somit nach dem Schulabschluss an deutlich weniger Bildungseinrichtungen eingeschrieben, die durch das ZSR erfasst werden (allgemein- und berufsbildende Schulen). Dabei handelt es sich um keinen Datenfehler, sondern ein empirisch bestätigtes Phänomen (Beicht 2015; Baas 2021).[1]

4.4. Datenqualität des ZSR und IVDS

Im folgenden Abschnitt werden insbesondere die Datenqualitätsdimensionen Genauigkeit, Vollständigkeit und Volatilität des ZSR und IVDS untersucht. Die Untersuchung erfolgt über den Vergleich der Angaben zwischen den Schuljahren sowie zwischen ZSR und IVDS. Die Datensätze wurden alle über die GUID verknüpft. Es gilt zu beachten, dass Datenfehler nur identifiziert werden können, falls Unterschiede zwischen den Jahren vorliegen. Datenfehler, die unverändert bleiben, können nicht als solche identifiziert werden.

4.4.1. Unterschiede zwischen den Schuljahren

Die Unterschiede zwischen den Schuljahren werden in Tabelle 4.3 dargestellt. Es zeigt sich deutlich, dass bei Migranten mehr Unterschiede zwischen den Merkmalen vorliegen. Im Vergleich zu Deutschen sind Migranten z. B. doppelt so häufig umgezogen (Unterschiede 100m-Gitter-ID und Geo-Koordinaten) und Vor- und Nachname unterscheiden sich etwa dreimal so häufig. Insbesondere der Vorname ist durch einen hohen Anteil fehlender Werte gekennzeichnet. Dies zeigt zunächst, dass bei Migranten eine insgesamt höhere Fehlerquote vorliegt. Die Ergebnisse von Menschen mit Migrationshintergrund ähneln zumeist den Ergebnissen von Deutschen.

Falls bei der Erfassung einer Person Vor- und Nachname nicht eindeutig getrennt werden können, existiert die Vorgabe, dass stattdessen der vollständige Name in das Feld des Nachnamens eingetragen werden soll. Der vollständige Name wird als *Blockname* bezeichnet (Koordinierungsstelle für IT-Standards 2022: 15). Fehlende Werte beim Vornamen sind auf diese Vorgabe zurückzuführen. Die wenigen fehlenden Werte beim Nachnamen lassen sich darüber erklären, dass die Regeln nicht korrekt befolgt und stattdessen der Blockname im Feld des Vornamens eingetragen wurde.

[1] Es können zudem Unterschiede bei der Bildungsbeteiligung der über 18-Jährigen nach Herkunft beobachtet werden. Die Erklärungen der Veränderungen sind jedoch nicht trivial. So liegen unterschiedliche Wirkungsmechanismen für die einzelnen Länder vor. Daher wird auf eine Darstellung der Ergebnisse im Rahmen dieser Arbeit verzichtet.

Ein möglicher Fehler bei Vor- und Nachnamen besteht zudem darin, dass die Felder miteinander vertauscht wurden. Dies war bei 36 Personen der Fall. Weitere 31 Personen enthielten die gleiche Ausprägung im Vor- und Nachnamen. Auch hier wurde somit die oben beschriebene Regel nicht richtig befolgt.

Die Tabelle zeigt auch, dass die Fehlerquoten für die QIDs nicht gleich verteilt sind. Insbesondere Geschlecht und Geburtsdatum weisen sehr wenige Fehler auf. So unterscheiden sich nur 0.2 % der vollständigen Geburtsdaten im ZSR. Davon weisen 52 Beobachtungen (0.03 %) zugleich eine Veränderung des Geburtstags, -monats und -jahrs auf.

Von weiterem Interesse ist die Zusammensetzung der Schüler, deren Daten von den Schulen stammen. Hierbei muss beachteten werden, dass für über 18-Jährige der Besuch einer schulischen Einrichtung freiwillig ist. Bei der Betrachtung dieser Personen fällt wie bereits gezeigt auf, dass 15.3 % der Migranten und 14 % der Deutschen im Datensatz über 18 Jahre alt sind, jedoch nur 2.3 % der Menschen mit Migrationshintergrund (Referenzjahr 2021). Dies führt dazu, dass ein höherer Anteil Menschen mit Migrationshintergrund durch die Schule gemeldet wurde. Um einen Zusammenhang zwischen der Datenquelle und der Herkunft zu beobachten, dürfen daher nur schulpflichtige Kinder betrachtet werden. Ferner zeigte die Betrachtung der imputierten Werte (Abbildung 4.1), dass die Anzahl durch die Schule gemeldeten Personen zwischen 16 und 18 Jahren ansteigt. Dieses Intervall muss daher getrennt betrachtet werden.

Für die Daten der 16- oder 17-Jährigen im Schuljahr 2021/22 kann beobachtet werden, dass 2.3 % der Deutschen, 1.6 % der Menschen mit Migrationshintergrund und 8.2 % der Migranten von der Schule gemeldet wurden. Bei den Daten der Kinder zwischen 6 und 15 Jahren beträgt der Anteil 0.6 % bei Deutschen, 0.7 % bei Menschen mit Migrationshintergrund und 1.6 % bei Migranten. Migranten werden somit deutlich häufiger von der Schule gemeldet. Es wurde absichtlich der imputierte Datensatz gewählt. Dadurch zeigen die Ergebnisse eindeutig, dass die Unterschiede auf unterschiedliche Wohnorte zurückzuführen sind. Migranten wohnen deutlich häufiger außerhalb Hamburgs und besuchen in Hamburg die Schule. Damit haben sie wahrscheinlich auch deutlich längere Schulwege als Deutsche oder Menschen mit Migrationshintergrund.

Ferner zeigen sich Unterschiede in den Fehlerquoten bei Schul- und Melderegisterdaten. Die Fehlerquoten nach Datenquelle und Herkunft für die Variablen des ZSR werden in Tabelle 4.4 dargestellt. Dabei werden Adressdaten zunächst ausgelassen. Diese werden in Abschnitt 4.4.2 behandelt.

Tabelle 4.4 zeigt, dass unterschiedliche Fehlerquoten nach Herkunft erhalten bleiben. Es liegen höhere Fehlerquoten vor, falls in mindestens einem Schuljahr die Daten von einer Schule stammen. Schulen machen somit deutlich mehr Fehler bei der Datenerfassung als Meldeämter. Dabei fallen besonders hohe Fehlerquoten bei Vor- und Nachname auf. Insbesondere Vornamen sind auffällig, da hier ein Großteil der Unterschiede auf echte Fehler und nicht auf Namensänderungen zurückzuführen sind. So liegen etwa 20-mal mehr Unterschiede im Vornamen bei den Schuldaten im Vergleich zum Melderegister vor. Besonders viele Fehler liegen bei Migranten vor.

Auch die Melderegisterdaten beinhalten Fehler. Dabei lassen sich auch unterschiedliche Fehlerquoten nach Herkunft beobachten. So unterscheiden sich 1.21 % der Vornamen von Migranten im Melderegister, jedoch nur 0.07 % der Vornamen von Deutschen. Für das Record-Linkage bedeuten die unterschiedlichen Fehlerquoten, dass auch auf Basis von Melderegisterdaten ein Linkage-Bias möglich ist. Dies hängt jedoch maßgeblich von der Art der Fehler ab. Eine detaillierte Aufschlüsslung der Fehler erfolgt deshalb in Abschnitt 4.5.

		Unterschiedliche Werte (%)				Wechsels zu fehlenden Wert (%)				Dauerhaft fehlender Wert (%)			
Variable	Quelle	Alle	Deut.	Mig.-h.	Mig.	Alle	Deut.	Mig.-h.	Mig.	Alle	Deut.	Mig.-h.	Mig.
Geschlecht	IVDS	0.03	0.01	0.02	0.18	0.00	0.00	0.00	0.00	7.54	8.00	2.85	13.09
Geburtsjahr	IVDS	0.04	0.02	0.01	0.21	0.00	0.00	0.00	0.00	7.54	8.00	2.85	13.09
Geburtsmonat	IVDS	0.06	0.01	0.03	0.35	0.00	0.00	0.00	0.00	7.54	8.00	2.85	13.09
Geburtsland	IVDS	0.11	0.00	0.25	0.54	0.00	0.00	0.01	0.00	7.56	8.01	2.86	13.19
1. Staatsbürgerschaft	IVDS	0.81	0.00	2.49	2.55	0.07	0.00	0.10	0.38	7.66	8.00	2.95	14.02
2. Staatsbürgerschaft	IVDS	0.06	0.00	0.26	0.01	1.80	0.00	7.79	1.77	85.49	100.00	30.59	97.47
Familiensprache	IVDS	1.11	0.67	2.45	1.38	0.05	0.03	0.05	0.12	7.59	8.03	2.92	13.26
RISE-Status	IVDS	9.89	9.12	10.56	13.16	0.61	0.60	0.56	0.77	9.15	9.96	3.73	13.90
Schultyp	IVDS	19.46	18.70	21.51	20.34	0.00	0.00	0.00	0.00	7.54	8.00	2.85	13.09
Geschlecht	ZSR	0.09	0.07	0.05	0.31	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Vorname	ZSR	1.10	0.93	0.86	2.55	0.10	0.00	0.14	0.63	0.22	0.00	0.26	1.44
Nachname	ZSR	0.74	0.43	0.76	2.50	0.00	0.00	0.00	0.03	0.01	0.00	0.00	0.06
Geburtstag	ZSR	0.10	0.05	0.03	0.52	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Geburtsmonat	ZSR	0.10	0.06	0.03	0.46	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Geburtsjahr	ZSR	0.08	0.05	0.03	0.32	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1. Staatsbürgerschaft	ZSR	1.11	0.00	1.75	6.43	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2. Staatsbürgerschaft	ZSR	0.04	0.00	0.16	0.08	2.34	0.00	8.63	4.96	82.48	100.00	19.21	91.56
Straße	ZSR	8.62	7.52	8.09	15.96	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
Hausnummer	ZSR	9.04	7.83	8.52	17.01	0.05	0.05	0.02	0.11	0.01	0.01	0.00	0.01
Postleitzahl	ZSR	6.66	5.57	6.28	13.70	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
100m-Gitter-ID	OSM	8.78	7.62	8.28	16.41	0.05	0.05	0.02	0.11	0.04	0.06	0.01	0.02
250m-Gitter-ID	OSM	8.51	7.37	8.01	16.02	0.05	0.05	0.02	0.11	0.04	0.06	0.01	0.02
500m-Gitter-ID	OSM	8.21	7.11	7.74	15.44	0.05	0.05	0.02	0.11	0.04	0.06	0.01	0.02
1km-Gitter-ID	OSM	7.74	6.64	7.31	14.92	0.05	0.05	0.02	0.11	0.04	0.06	0.01	0.02
Koordinate	OSM	9.09	7.90	8.57	16.99	0.05	0.05	0.02	0.11	0.04	0.06	0.01	0.02

Tabelle 4.3.: Unterschiede zwischen den Records in beiden Schuljahren, aufgeteilt nach Herkunft (Deut. = Deutsch, Mig.-h. = Migrationshintergrund, Mig. = Migrant). Der Block "unterschiedliche Werte" beschreibt den Anteil beobachteter Werte, die sich zwischen den Schuljahren unterscheiden. Der Anteil Änderungen von einem beobachteten zu einem fehlenden Wert wird im Block "Wechsel zu fehlenden Wert" erfasst (ohne Berücksichtigung der Reihenfolge). Im Block "Dauerhaft fehlende Werte" wird der Anteil Records einer Variable beschrieben, bei denen in beiden Schuljahren ein fehlender Wert vorliegt.

		Unterschiedliche Werte (%)				Wechsels zu fehlenden Wert (%)				Dauerhaft fehlender Wert (%)			
Variable	Quelle	Alle	Deut.	Mig.-h.	Mig.	Alle	Deut.	Mig.-h.	Mig.	Alle	Deut.	Mig.-h.	Mig.
Geschlecht	Schule	0.58	0.51	0.57	0.88	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Meld.	0.02	0.00	0.01	0.18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Beide	0.35	0.31	0.28	0.63	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Vorname	Schule	4.64	4.34	5.48	5.65	0.03	0.00	0.00	0.15	0.00	0.00	0.00	0.00
	Meld.	0.23	0.07	0.27	1.21	0.08	0.00	0.10	0.56	0.26	0.00	0.29	1.78
	Beide	6.75	6.92	5.09	8.77	0.42	0.02	0.52	1.76	0.12	0.00	0.11	0.59
Nachname	Schule	1.86	1.36	1.83	4.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Meld.	0.51	0.30	0.56	1.78	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.07
	Beide	2.02	0.87	2.28	6.02	0.03	0.00	0.00	0.18	0.00	0.00	0.00	0.00
Geburtstag	Schule	0.50	0.39	0.57	0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Meld.	0.05	0.00	0.00	0.43	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Beide	0.26	0.19	0.17	0.72	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Geburtsmonat	Schule	0.62	0.51	0.91	0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Meld.	0.04	0.01	0.00	0.31	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Beide	0.29	0.19	0.14	0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Geburtsjahr	Schule	0.55	0.49	0.46	0.84	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Meld.	0.02	0.00	0.00	0.18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Beide	0.24	0.12	0.19	0.77	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1. Staatsbürgerschaft	Schule	1.32	0.00	3.08	6.41	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Meld.	0.95	0.00	1.48	5.93	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Beide	2.52	0.00	3.83	10.13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2. Staatsbürgerschaft	Schule	0.07	0.00	0.23	0.34	2.06	0.00	31.51	1.07	95.27	100.00	26.37	97.94
	Meld.	0.03	0.00	0.12	0.03	1.57	0.00	4.61	5.46	81.79	100.00	19.79	90.44
	Beide	0.17	0.00	0.55	0.18	10.82	0.00	39.44	5.79	76.71	100.00	12.25	92.27

Tabelle 4.4.: Unterschiede zwischen den Records im ZSR, aufgeteilt nach Herkunft (Deut. = Deutsch, Mig.-h. = Migrationshintergrund, Mig. = Migrant) und Datenquelle beider Schuljahre (Schule, Melderegister oder beide). Der Block "unterschiedliche Werte" beschreibt den Anteil beobachteter Ausprägungen, die sich zwischen den Schuljahren unterscheiden. Der Anteil Änderungen von einem beobachteten zu einem fehlenden Wert wird im Block "Wechsel zu fehlenden Wert" erfasst (ohne Berücksichtigung der Reihenfolge). Im Block "Dauerhaft fehlende Werte" wird der Anteil Records einer Variable beschrieben, bei denen in beiden Schuljahren ein fehlender Wert vorliegt.

Anhand des Vornamensfeldes ist zudem erkennbar, dass im Melderegister Blocknamen verwendet werden, währenddessen die Schulen meist den Namen aufteilen. Dies zeigt sich durch den Anteil dauerhaft fehlender Werte sowie durch den hohen Anteil Wechsel zu einem fehlenden Wert beim Vornamen.

Weiterhin auffällig sind Unterschiede in der zweiten Staatsbürgerschaft. Ein häufiger Wechsel zu einem fehlenden Wert deutet darauf hin, dass die zweite Staatsbürgerschaft in den Schuldaten häufig nicht erfasst wird. Dies zeigt sich besonders beim Wechsel der Datenquellen. Bei 39.4 % der Menschen mit Migrationshintergrund fehlt die zweite Staatsbürgerschaft, wenn die Daten in einem Schuljahr von der Schule und im anderen Schuljahr vom Melderegister stammen. Hieraus kann gefolgert werden, dass unterschiedliche Erfassungspraktiken bei den Schulen existieren. Der geringere Anteil Wechsel bei Migranten lässt sich auf die Definition von Migranten zurückführen. So besitzen 8.4 % der Migranten und 80.8 % der Menschen mit Migrationshintergrund eine zweite Staatsbürgerschaft.

Der Wechsel der ersten Staatsbürgerschaft ist bei Migranten, insbesondere beim Wechsel der Datenquelle, mit 10.1 % auffällig. Es muss beachtet werden, dass nach § 37 StAG (Staatsangehörigkeitsgesetz) eine eigenständige Einbürgerung erst ab dem 16. Lebensjahr möglich ist. Angesichts dessen erscheint es als möglich, dass der Wechsel der Datenquelle mit Einbürgerungen (Wechsel der ersten Staatsbürgerschaft) korreliert.

Die Tabellen 4.3 und 4.4 zeigen, dass insgesamt sehr wenige fehlende Werte vorliegen. Der höhere Anteil fehlender Werte im IVDS ist dadurch zu erklären, dass für diese Personen kein Eintrag aus dem IVDS vorliegt. Der Großteil der Records des ZSR und IVDS ist vollständig und für beide Schuljahre fast identisch. Unter Ausschluss der zweiten Staatsbürgerschaft liegen für das Schuljahr 2021/22 bei 99.7 % der Records des ZSR und 97 % der Records des IVDS keine fehlenden Werte vor. Bei allen Records des ZSR fehlt maximal ein Wert, wobei hier die Adresse als ein Merkmal zusammengefasst wird. 2.9 % der Records im IVDS haben einen fehlenden Wert. Für 35 Beobachtungen liegen zwei fehlende Werte vor und für 2 Beobachtungen drei fehlende Werte. Bei 91.3 % der Records mit fehlenden Werten im IVDS ist der RISE-Status nicht angegeben. Wiederum 80.4 % der fehlenden RISE-Werte können darauf zurückgeführt werden, dass die betroffene Person nicht in Hamburg wohnte. Angesichts dieses niedrigen Anteils fehlender Werte wird auf eine weitere Analyse der fehlenden Werte im Folgenden verzichtet.

4.4.2. Umzüge

Eine zentrale Fehlerursache ist höchstwahrscheinlich eine Neuerfassung der Daten. Dies wurde auch in der Simulation angenommen (Schnell 2022). Da eine Neuerfassung oder Änderung der Daten nach einem Umzug notwendig ist, werden diese im Folgenden untersucht. Eine Betrachtung von Umzügen ist darüber hinaus für eine Optimierung des Record-Linkage von Interesse. So kann beispielsweise die geografische Distanz zum Vergleich von Records verwendet werden.

Umzüge werden definiert als eine Veränderung der Geo-Koordinate im Untersuchungszeitraum. Insbesondere für die altersbedingte Umzugswahrscheinlichkeit muss bedacht werden, dass es sich um die Wahrscheinlichkeit handelt, im Untersuchungszeitraum umzuziehen. So beträgt z. B. die Umzugswahrscheinlichkeit der 18-Jährigen 15.4 %. Diese besagt, dass 15.4 % der Menschen, die 2021 18 Jahre alt waren, in den folgenden zwei Jahren mindestens einmal umgezogen sind.

Es muss zudem beachtet werden, dass Hamburg eine Großstadt ist und daher eine besondere Wohnsituation aufweist. Auch Hamburg ist durch einen Wohnungsmangel geprägt, der sich bundesweit in vielen Großstädten zeigt (NDR 2023). Beides kann zu Umzugsmustern führen, die beispielsweise mit

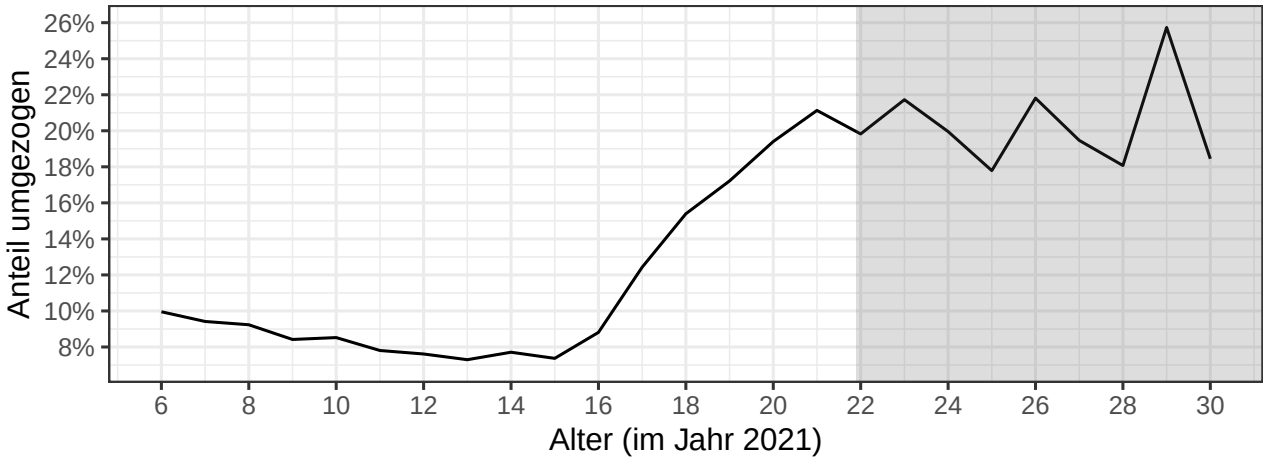


Abbildung 4.5.: Umzugswahrscheinlichkeit nach Alter in Jahren. Der grau hinterlegte Bereich kennzeichnet alle Jahre mit weniger als 1000 Beobachtungen.

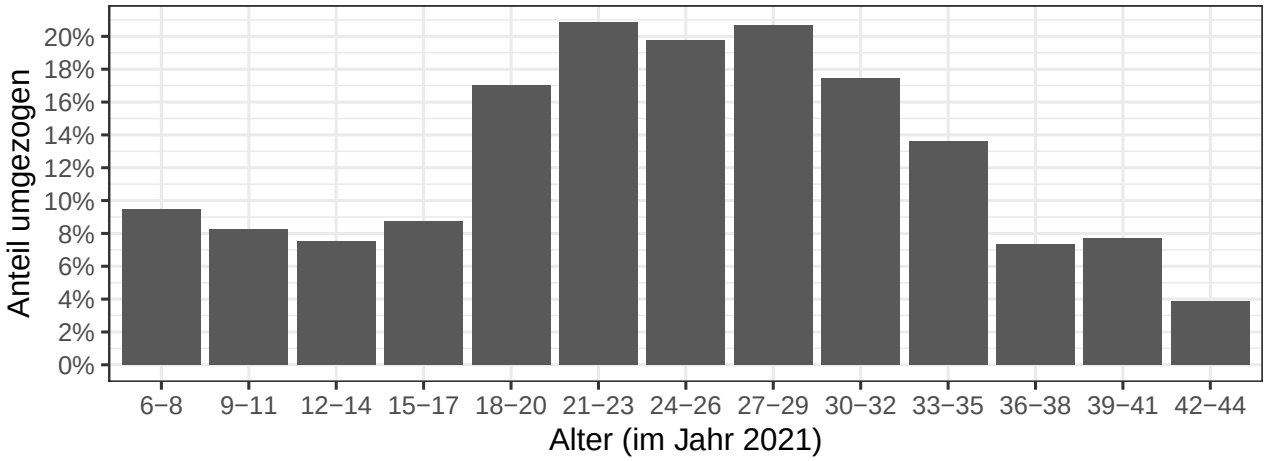


Abbildung 4.6.: Umzugswahrscheinlichkeit nach Altersgruppen.

ländlichen Regionen nicht vergleichbar sind. So sind die Distanzen wahrscheinlich deutlich kleiner und Umzüge werden aufgrund des fehlend einer passenden Wohnung zeitlich häufiger verschoben.

Für die Darstellung der Umzugswahrscheinlichkeit muss beachtet werden, dass die Anzahl Beobachtungen nach Ende der Schulpflicht stark abnimmt. Daher liegen in der Umzugswahrscheinlichkeit älterer Personen stärkere zufällige Schwankungen vor, die nicht interpretiert werden können. Die Darstellungen wurden daher rechts zensiert. Abbildung 4.5 stellt zunächst die altersbedingte Umzugswahrscheinlichkeit für die gesamte Population zwischen 6 und 30 Jahre dar, wobei ab dem 23. Lebensjahr bereits starke Schwankungen zu beobachten sind. Aus diesem Grund präsentiert Abbildung 4.6 die Umzugswahrscheinlichkeit nach Altersgruppen. Abbildung 4.7 stellt die altersbedingte Umzugswahrscheinlichkeit nach Herkunft dar. Da insbesondere Menschen mit Migrationshintergrund nach Vollendung der Schulzeit fehlen, ist diese Darstellung nur in der Altersspanne von 6 bis 17 Jahren sinnvoll interpretierbar.

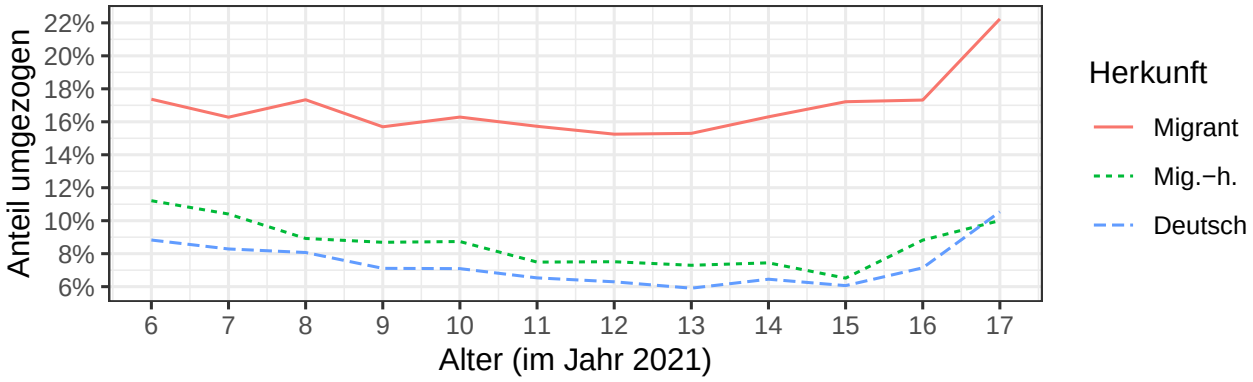


Abbildung 4.7.: Umzugswahrscheinlichkeit nach Alter und Herkunft.

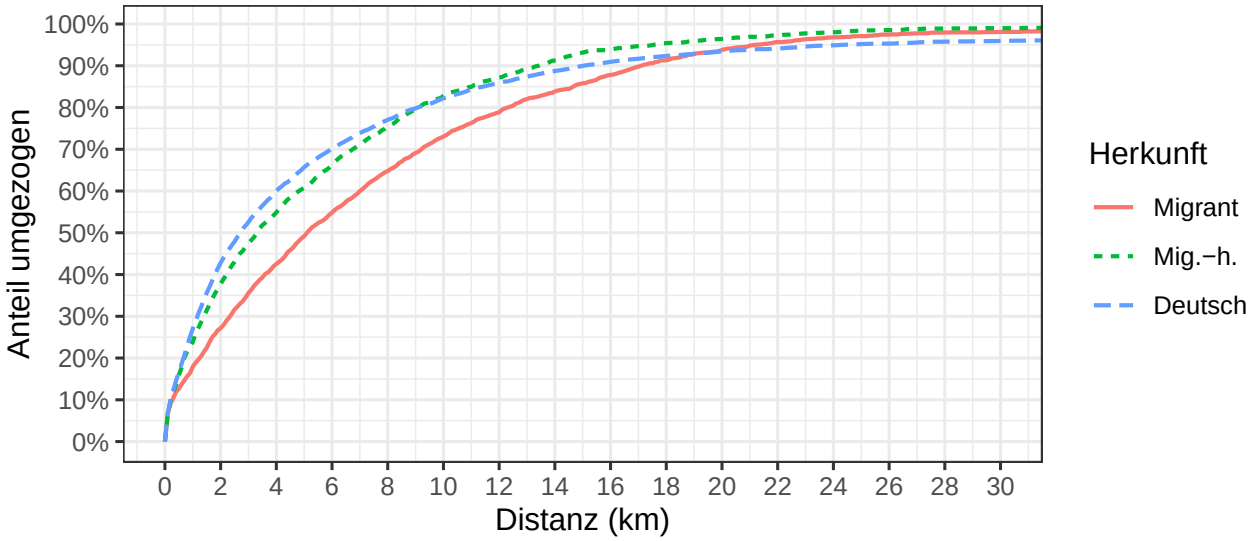


Abbildung 4.8.: Kumulierter Anteil Umzüge bis 30 km nach Distanz und Herkunft.

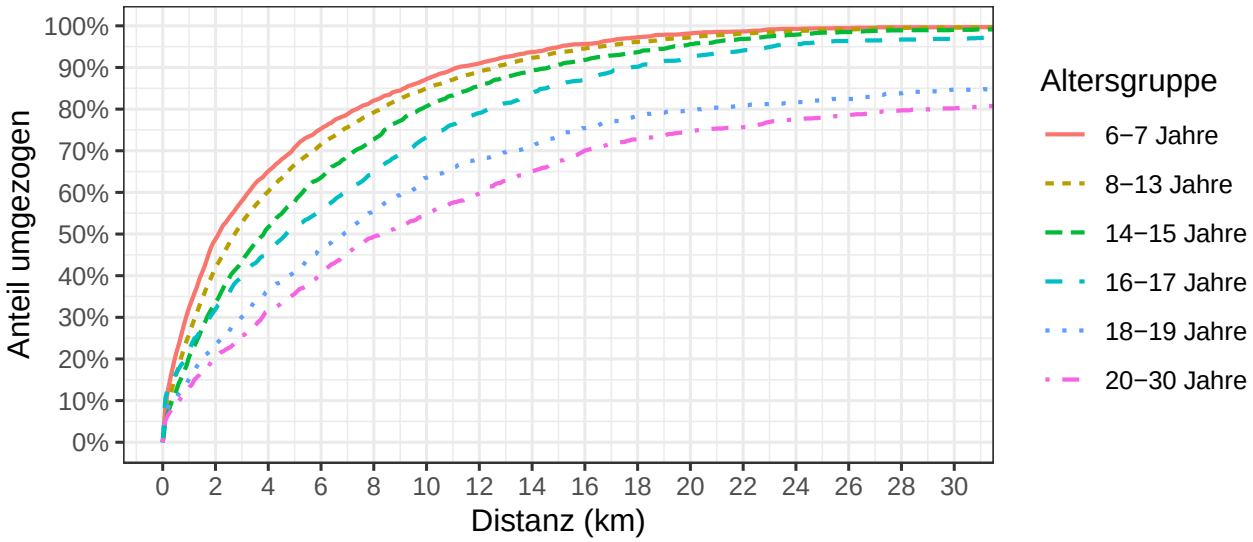


Abbildung 4.9.: Kumulierter Anteil Umzüge bis 30 km nach Distanz und Alter (gruppiert).

Zunächst zeigt Abbildung 4.5, dass die Umzugswahrscheinlichkeit bis zum 15. Lebensjahr linear sinkt und dann mit dem Beginn einer Ausbildung und der Volljährigkeit stark ansteigt.[1] Die durchschnittliche Umzugswahrscheinlichkeit für Schüler beträgt 8.5 % (Kinder jünger als 18 Jahre). Die Umzugswahrscheinlichkeit für Schüler in der Simulation betrug 7.8 % und wurde somit gut geschätzt (Schnell 2022; Weiand 2022).[2]

Abbildung 4.6 zeigt, dass die Umzugswahrscheinlichkeit zwischen 18 und 35 Jahren mit bis zu 20.8 % deutlich erhöht ist. Danach sinkt die Wahrscheinlichkeit auf ein deutlich niedrigeres Niveau von etwa 7.5 %. Es muss beachtet werden, dass die Gruppe der über 18-Jährigen stark selektiv ist und mit den Ergebnissen der Simulation nicht direkt verglichen werden kann. Dennoch weist die Abbildung darauf hin, dass insbesondere im Altersbereich weiterführender Bildung mit einer erhöhten Anzahl Umzügen zurechnen ist.

Die altersbedingte Umzugswahrscheinlichkeit nach Herkunft (Abbildung 4.7) zeigt, dass Migranten eine deutlich höhere Umzugswahrscheinlichkeit haben. Die mittlere Wahrscheinlichkeit für Schüler ist mit 16.5 % bei Migranten mehr als doppelt so groß wie für Deutsche (7.2 %). Dies wurde in der Simulation nicht berücksichtigt. Die höhere Umzugswahrscheinlichkeit kann zu einer noch schlechteren Linkage-Qualität bei Migranten führen, als in der Simulation veranschlagt wurde.

Neben einer höheren Umzugswahrscheinlichkeit ist die Umzugsdistanz bei Migranten zumeist auch höher. Abbildung 4.8 zeigt den kumulierten Anteil umgezogener Personen nach Entfernung der Wohnorte. So ziehen 50 % der Deutschen in einem Radius von 2.8 km um, während 50 % der Migranten in einem Radius von 5.2 km umziehen. Ein Vergleich anhand der räumlichen Nähe ist für Migranten somit auch ineffektiver. Zur Lesbarkeit wurde die Umzugsdistanz nach 30 km rechts zensiert. Zum Vergleich: Die geografische Ausdehnung des Stadtgebiets von Hamburg (ohne Inseln) beträgt etwa 40 km in Nord-Süd und Ost-West-Richtung.

Als Letztes stellt Abbildung 4.9 die Umzugsdistanz nach Altersgruppen dar. Zur übersichtlichen grafischen Darstellung ist die Abbildung auf wenige Altersgruppen beschränkt. Dies ist das Grundschulalter (6–7 Jahre), die weiterführende Schule (8–13 Jahre), die Mittleren Reife (14–15 Jahre), das Abiturs bzw. die Berufsausbildung (16–17 Jahre) sowie das Schulende (18–19 Jahre und 20–30 Jahre). Das Referenzjahr ist 2021.[3]

Die Ergebnisse zeigen, dass besonders zwischen zwei Gruppen unterschieden werden kann. Bis zum 18. Lebensjahr ziehen Menschen über kleinere Distanzen um, wobei sich die Distanz mit wachsendem Alter leicht erhöhen. Fast alle ziehen in einem 30km-Radius um. Mit Erreichen der Volljährigkeit ziehen die Menschen nicht nur häufiger, sondern auch weiter weg. 20 % der 20- bis 30-Jährigen sind mehr als 30 km weit umgezogen. Die Steigung der Geraden für die 18–19-Jährigen und 20–30-Jährigen bleibt nach 30 km in etwa konstant und endet bei einer Umzugsdistanz von 656 km (18–19 Jahre) und 668 km (20–30 Jahre).

[1] Der linear sinkende Trend lässt sich handlungstheoretisch wahrscheinlich dadurch erklären, dass mit zunehmendem Alter der Kinder mehr Eltern eine gefestigte Wohnsituation haben. Gleichzeitig steigt der Nutzen, einen ggf. notwendigen Umzug zeitlich zu verschieben, da eine künftige Änderung der Wohnsituation durch Auszug der Kinder wahrscheinlicher wird.

[2] In der Simulation wurde eine Umzugswahrscheinlichkeit von 4 % für ein Jahr angenommen (Schnell 2022). Unter Berücksichtigung von Mehrfachumzügen somit die Umzugswahrscheinlichkeit nach zwei Jahren $p = 0.04 \cdot 0.96 \cdot 2 + 0.04 \cdot 0.04 = 0.0784$.

[3] Bei der Gruppeneinteilung muss der Referenzzeitraum beachtet werden. So beinhaltet z. B. die Gruppe der Kinder im Grundschulalter alle Kinder, die zum Ende der Referenzzeit 9 Jahre alt sind und sich daher immer noch in der Grundschule befinden.

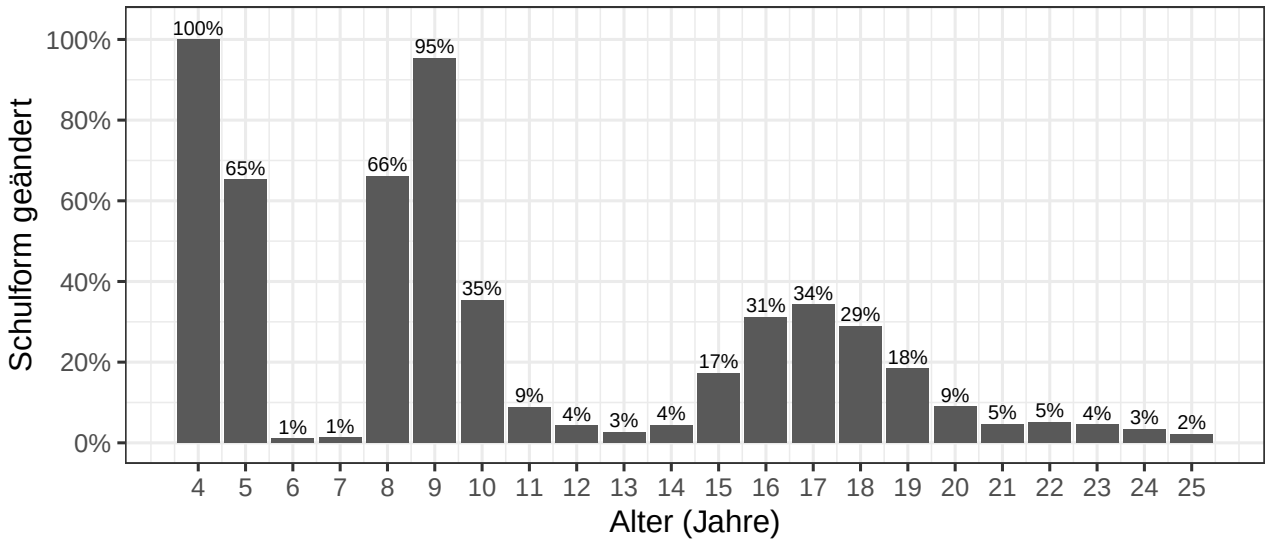


Abbildung 4.10.: Anteil Personen nach Alter, deren Schulform sich geändert hat.

4.4.3. Fehlerursachen

Die bisherige Aufführung weist darauf hin, dass es mindestens vier verschiedene Ereignisse gibt, die einen Fehler in den Daten verursachen können. Zunächst sind dies Umzüge und Schulwechsel. Diese wurden auch in der Simulation als Fehlerursache gesehen (Schnell 2022). Darüber hinaus wurde bereits gezeigt, dass viele Fehler mit der Veränderung der Datenquelle einhergehen. Eine weitere mögliche Ursache für Fehler ist die Korrektur von unvollständigen Angaben. Falls ein Migrant das Geburtsdatum 1.1. besitzt, ist diese besonders bei entsprechenden Geburtsdaten möglich. Da Schulwechsel und das Geburtsdatum 1.1. bisher nicht genauer betrachtet wurden, folgt zunächst eine Beschreibung dieser beiden Phänomene.

Abbildung 4.10 stellt die Anteile der Personen nach Alter dar, bei denen sich die Schulform geändert hat. Gegeben der Datengrundlage ist dies die bestmögliche Annäherung an die tatsächlich erfolgten Schulwechsel. So gilt es zu beachten, dass ein Wechsel der Schule ohne Wechsel der Schulform nicht identifizierbar ist. Es kann jedoch angenommen werden, dass ein Großteil der Schulwechsel auch mit einem Wechsel der Schulform einhergeht.

Abbildung 4.10 zeigt deutlich, dass zunächst alle Vorschulkinder in die Grundschule wechseln. Weiter wechseln fast alle Kinder, die sich in der dritten und vierten Klasse befinden, auf eine weiterführende Schule. Es gilt zu beachten, dass das Alter zum Stichtag für die Einschulung und das Alter im Datensatz nicht übereinstimmen. Folglich verteilen sich die Wechsel von der Grundschule zur weiterführenden Schule besonders auf Kinder zwischen 8 und 10 Jahren. Die Wechsel von der weiterführenden Schule zu einer anderen Schulform erfolgen deutlich weiter gestreut, in einem Alter zwischen 15 und 20 Jahren. Da nur die Daten des IVDS für allgemeinbildende Schulen vorliegen, sind die letzten Wechsel zumeist ein Wechsel zu einer unbekannten Schulform. In den meisten Fällen sind dies wahrscheinlich Wechsel zu einer Berufsschule, da Universitäten und Fachhochschule nicht im ZSR erfasst werden (Abschnitt 4.1.1).

Unter dem Geburtsdatum 1.1. wird im Folgenden ein Migrant verstanden, bei dem in mindestens einem der beiden Datensätze der 1. Januar als Geburtsdatum eingetragen wurde. Es werden ausschließlich Migranten betrachtet, für die eine Beobachtung in beiden Schuljahren vorliegen. Unter dieser Definition haben 6.1 % der Migranten das Geburtsdatum 1.1. Durchaus können hierunter auch

		Anteil Veränderungen (in %)					
Variable	Quelle	Umzug	Schulwechsel	Datenquelle	1.1.	Gesamt	n
Geschlecht	IVDS	19.05	34.92		6.35	50.79	63
Geburtsjahr	IVDS	16.42	41.79		11.94	59.70	67
Geburtsmonat	IVDS	18.63	26.47		18.63	50.00	102
Geburtsland	IVDS	18.66	56.46		1.44	67.94	209
1. Staatsbürgerschaft	IVDS	11.89	27.45		2.45	37.64	1472
2. Staatsbürgerschaft	IVDS	8.91	29.70		0.00	35.64	101
Familiensprache	IVDS	13.05	44.12		0.89	51.80	2031
Mittelwert	IVDS	15.23	37.27		5.95	50.50	
Geschlecht	ZSR	37.50	25.00	16.67	10.42	72.92	48
Vorname	ZSR	17.96	18.41	40.87	2.10	65.27	668
Nachname	ZSR	13.63	25.05	13.96	1.65	45.16	910
Geburtstag	ZSR	24.72	19.10	17.98	16.85	61.80	89
Geburtsmonat	ZSR	21.43	22.86	14.29	21.43	55.71	70
Geburtsjahr	ZSR	19.51	31.71	24.39	14.63	63.41	41
1. Staatsbürgerschaft	ZSR	11.35	23.51	10.48	5.27	44.60	1612
2. Staatsbürgerschaft	ZSR	16.67	26.67	30.00	0.00	61.67	60
Mittelwert	ZSR	20.35	24.04	21.08	9.04	58.82	

Tabelle 4.5.: Anteil Veränderungen, die in Verbindung mit einem Ereignis aufgetreten sind, sowie die absolute Anzahl Veränderungen (n). Es wurden nur Fälle eingeschlossen, für die Daten aus dem ZSR und IVDS für beide Schuljahre vorlagen.

Personen fallen, die tatsächlich am 1. Januar geboren wurden. Unter Abzug der geschätzten Anzahl Migranten, die an diesem Tag geboren sein könnten, ergibt sich weiterhin ein Anteil von 5.8 % Migranten, denen kein korrektes Geburtsdatum zugewiesen wurde.[1]

Tabelle 4.5 stellt den Anteil Veränderungen dar, die in Zusammenhang mit einem Ereignis stehen (Schulwechsel, Umzug, Veränderung der Datenquelle und Geburtsdatum 1.1.) sowie den Anteil Veränderungen, die insgesamt erklärt werden können. Da die Information für den Wechsel der Schulform aus dem IVDS stammt, wurden nur Personen eingeschlossen, für die eine Beobachtung aus dem IVDS und ZSR für beide Schuljahre vorlag. Betrachtet werden daher nur Schüler von allgemeinbildenden Schulen. Die Veränderung der Datenquelle wird zudem nur für das ZSR dargestellt, da das Ereignis keinen Einfluss auf den IVDS hat. Überdies sind die Anteile für die einzelnen Ereignisse nicht unabhängig. So sind mehrere Ereignisse simultan möglich. In diesem Fall wurde eine erklärte Veränderung zu allen betroffenen Ereignissen gerechnet. Da der Anteil Veränderungen mit simultanen Ereignissen gering ist, wurde aus Zwecken der Darstellbarkeit und Interpretierbarkeit auf eine getrennte Darstellung verzichtet. Dies ist auch möglich, da alle Ereignisse (Schulwechsel, Umzug, Veränderung der Datenquelle und Geburtsdatum 1.1.) nur sehr schwach miteinander korrelieren ($\phi < 0.09$).

Im Mittel stehen nur etwa 59 % der Veränderungen im ZSR und 51 % der Veränderungen im IVDS mit einem der vier Ereignisse in Zusammenhang. Die Ereignisse können somit nicht alle Veränderung erklären. Insbesondere eine Veränderung des Nachnamens und der ersten Staatsbürgerschaft kann zu einem geringen Teil durch die vier Ereignisse erklärt werden. Jedoch existieren weitere Erklärungen für die Veränderungen dieser beider Merkmale. So kann sich der Nachname durch Annahme oder Ablegen eines Ehenamens auch bei Kindern ändern. Für die Namensänderung eines Kindes gibt es drei Möglichkeiten:

[1] Für die Berechnung wird angenommen, dass die Geburtsdaten von Migranten gleichverteilt sind. Berechnungen, die auf der empirischen Geburtsdatenverteilung aufbauen, kommen zu fast identischen Ergebnissen.

1. Einbenennung (§ 1618 BGB): Ein Kind, das außerhalb der Ehe geboren wurde, kann unter Zustimmung beider Elternteile und des Kindes den Ehenamen oder einen Doppelnamen annehmen.
2. Namensänderung (§§ 3–4 NamÄndG): Unter Vorlage wichtiger Gründe kann der Name geändert werden (z. B. bei einer Scheidung). Bei Änderung des Nachnamens erstreckt sich diese Änderung auch auf die Kinder der Person, falls diese das Sorgerecht hat.
3. Adoption (§§ 1754, 1757 BGB): Bei einer Adoption erhält das Kind die rechtliche Stellung eines leiblichen Kindes der Adoptiveltern. Damit erhält das Kind auch den Familiennamen der Adoptiveltern. Eine gleichzeitige Änderung des Vornamens ist auf Antrag möglich. Eine zwangsläufige Änderung des Nachnamens erfolgt auch bei einer Erwachsenenadoption (§ 1767 BGB).[1]

Es existieren keine Zahlen über die Häufigkeit der Anwendung dieser Gesetze in Deutschland. In Hamburg wird eine Einbenennung und die durch eine Adoption folgende Namensänderung durch die Standesämter durchgeführt. Auf Anfrage teilten zwei der sieben Standesämter in Hamburg mit, dass sie etwa 10 Einbenennungen im Jahr durchführen. Häufig würde dabei ein Doppelname gewählt. Eine Kinderadoption ist aufgrund der hohen rechtlichen Hürden ein seltener Fall und erfolgt in den allermeisten Fällen im Kleinkindalter. Aufgrund dieser Hürden erfolgt häufiger eine Erwachsenenadoption, welche oft gleich mit Beginn der Volljährigkeit beantragt wird. Das Standesamt Bergedorf schätzte, dass etwa zwei Drittel aller Adoptionsanträge Erwachsenenadoptionen sind.

Für eine Namensänderung ist in Hamburg die Behörde für Inneres und Sport zuständig. Auf Anfrage konnte die Behörden keine Zahlen zu Namensänderungen von Kindern nennen. Es wurde jedoch darauf hingewiesen, dass dies “oft” passiert. Es ist somit plausibel, dass ein Großteil der gefundenen Veränderungen des Nachnamens auf einer rechtlich gültigen Änderung basieren.

Besonders die erste Staatsbürgerschaft weist deutlich mehr Veränderungen auf als andere Zahlenwerte. Diese können wahrscheinlich zu großen Teilen durch Einbürgerungen und nicht durch einen Datenfehler erklärt werden. Hierbei muss beachtet werden, dass bei der Einbürgerung eines Elternteils auch minderjährige Kinder eingebürgert werden können (§§ 9–10 StAG). Dementsprechend erscheint eine Einbürgerung als plausibel.

Umzüge, Schulwechsel und die Veränderung der Datenquelle wirken sich relativ gleichmäßig auf alle Variablen aus. Das Vorliegen des Geburtsdatums 1.1. bei Migranten steht im Wesentlichen nur mit einer Veränderung des Geburtsdatums in einem Zusammenhang. Auffällig ist zudem der häufige Wechsel des Vornamens bei Veränderung der Datenquelle. Dies deutet erneut darauf hin, dass Schulen den Vornamen häufig inkorrekt erfassen.

4.5. Fehleranalyse

In Abschnitt 4.4.1 wurden bereits die Unterschiede zwischen den verschiedenen Variablen aufgeführt. Diese Unterschiede lassen sich einerseits auf Tippfehler und anderseits auf eine unterschiedliche Angabe von Namensbestandteilen zurückführen. Beides wird im Folgenden gezeigt und analysiert.

Obwohl Tippfehler eine häufige Ursache für Fehler in Datensätzen darstellen, finden sich in der Literatur nur wenige Tippfehleranalysen (Kukich 1992). Daher erfolgt zunächst eine Einführung zu theoretischen Grundlagen der Tippfehleranalyse sowie zur implementierten Vorgehensweise. Anschließend werden die Fehler nach Fehlerart vorgestellt. Betrachtet werden die Merkmale Geburtstag, -monat,

[1] Das Bundesverfassungsgericht hat die zwangsläufige Namensänderung bei einer Erwachsenenadoption als verfassungswidrig eingestuft. Mit der Reform des Namensrechts ab dem 1. Mai 2025 soll eine zwangsläufige Namensänderung nicht mehr erfolgen (Bundesministerium für Justiz 2023).

-jahr, Postleitzahl, Vorname und Nachname sowie der Blockname. In Abschnitt 4.4.1 wurde bereits gezeigt, dass einige Fehler bei Namen durch die Verwendung von Blocknamen zustande kommen. Zudem zeigte sich, dass nicht alle Namensbestandteile richtig verortet wurden. Daher wird die Betrachtung des Blocknamens als sinnvoll erachtet. Der Blockname wurde für alle Personen durch Zusammenfassen von Vor- und Nachname mit trennendem Leerzeichen gebildet.

4.5.1. Grundlagen zur Analyse von Tippfehlern

Die Ausprägung einer Variable kann aus mehreren einzelnen Bestandteilen zusammengesetzt sein. Dies gilt insbesondere für Namen. So können Vornamen aus z. B. einem Erst- und Zweitnamen bestehen oder ein Nachname aus einem Doppelnamen. Die einzelnen Bestandteile werden im Folgenden als *Token* bezeichnet. Ein Token ist eine zusammenhängende Zeichenkette, die von einem Leerzeichen oder Bindestrich begrenzt wird. Der Name ANNA-MARIA MÜLLER besteht somit z. B. aus drei Token (ANNA, MARIA und MÜLLER).

Bei Analyse von Tippfehlern muss zudem zwischen einer Analyse von *Wörtern* und *Nicht-Wörtern* unterschieden werden. Ein Wort ist in einem Wörterbuch zu finden. In dem Wörterbuch wird eine eindeutige Schreibweise des Wortes aufgeführt. Eine Überprüfung der Korrektheit des Wortes kann daher anhand des Wörterbuchs erfolgen. Schwieriger und deutlich seltener analysiert sind Fehler in Nicht-Wörtern. Diese stehen nicht in einem Wörterbuch und können daher nicht auf Korrektheit überprüft werden. Hierzu zählen insbesondere Namen und Zahlen (Kukich 1992).

Fehler in Nicht-Wörtern können nur durch den Vergleich mit dem korrekten Wert vollständig analysiert werden (ground truth data). Die korrekten Werte liegen für die Daten nicht vor, sodass nur eine Annäherung anhand der Veränderungen zwischen den Jahren erfolgen kann.

Ein Problem bei der Analyse von Tippfehlern bei Wörtern besteht in der Größe des Wörterbuchs. So zeigt Peterson (1986), dass bei wachsender Größe des Wörterbuchs die Zahl der unerkannten Tippfehler zunimmt. So steigt die Wahrscheinlichkeit, dass durch einen Tippfehler ein anderes korrektes Wort gebildet wird. Dieses Problem lässt sich auch auf Fehler in Zahlenwerten anwenden. So können Zahlenwerte häufig auf eine gültige Menge reduziert werden. Fehler, die einen Wert innerhalb der gültigen Menge erzeugen, bleiben jedoch unerkannt. Der Fehler ist somit syntaktisch genau (Abschnitt 3.3).

Zur Klassifizierung von Fehlern wird im Folgenden die Damerau-Levenshtein-Metrik verwendet (Damerau 1964; Levenshtein 1966). Ausgehend von einem Startwort S und einem Zielwort Z definiert die Metrik vier Operationen, die ausgeführt werden können, um S in Z zu überführen:

- Ersetzung: Ein Zeichen muss durch ein anderes Zeichen ersetzt werden (z. B. S=NIKO und Z=NICO).
- Löschung: Ein Zeichen muss entfernt werden (z. B. S=PETER und Z=PETE).
- Einfügung: Ein Zeichen muss eingefügt werden (z. B. S=ANA und Z=ANNA).
- Vertauschung: Zwei benachbarte Zeichen müssen vertauscht werden (z. B. S=MARAI und Z=MARIA).

Damerau (1964) stellte fest, dass 80 % der von ihm analysierten Tippfehler in diese Metrik fallen. Die minimale Anzahl Operationen, die es benötigt, um S in Z zu überführen, wird als Damerau-Levenshtein-Distanz bezeichnet. Die oben aufgeführten Beispiele haben alle eine Damerau-Levenshtein-Distanz von 1.

Die Berechnung der Damerau-Levenshtein-Distanz erfolgt über eine Distanzmatrix. Durch rückwärts durchlaufen (traversieren) der vorher gebildeten Distanzmatrix ist es möglich, die einzelnen notwendi-

		D	O	O	H	T	E	R
	0	1	2	3	4	5	6	7
D	1	0	1	2	3	4	5	6
O	2	1	0	1	2	3	4	5
R	3	2	1	1	2	3	4	4
O	4	3	2	1	2	3	4	5
T	5	4	3	2	2	2	3	4
H	6	5	4	3	2	2	3	4
E	7	6	5	4	3	3	2	3
A	8	7	6	5	4	4	3	3

Abbildung 4.11.: Distanzmatrix und Traversion vom Startwort DOROTHEA zum Zielwort DOOHTER.

gen Operationen zur optimalen Überführung der Zeichenketten zu identifizieren. Die Dimensionen der Matrix M sind das Startwort S mit der Länge l_S und Zielwort Z mit der Länge l_Z. Beiden Wörtern wird am Anfang ein fehlendes Zeichen hinzugefügt, sodass M die Größe $(l_S + 1) \times (l_Z + 1)$ besitzt. Ein Teilausschnitt eines Wortes wird im Folgenden mit einem Suffix gekennzeichnet. S_i bedeutet z. B. das Startwort bis zum i-ten Buchstaben.

In jeder Zelle der Matrix $M_{i,j}$ wird die Distanz von S_i zu Z_j angegeben, wobei $i \in [0, l_S]$ und $j \in [0, l_Z]$. Aus diesem Grund wird die erste Spalte und erste Zeile der Matrix zu Beginn mit aufsteigenden Zahlen gefüllt, beginnend bei 0. So benötigt es nur Einfügungen oder Löschungen, um von einer leeren Zeichenkette zum Start- oder Zielwort zu gelangen. Die Distanzmatrix wird danach iterativ gebildet.

Beginnend bei der ersten freien Zelle $M_{1,1}$ beträgt der Wert der Zelle M_{i_j}:

$$M_{i,j} = \min \begin{cases} M_{i-1,j-1}, & \text{falls } S_i = Z_j \text{ (keine Operation)} \\ M_{i-1,j-1} + 1, & \text{falls } S_i \neq Z_j \text{ (Ersetzung)} \\ M_{i-1,j} + 1 & \text{(Löschung)} \\ M_{i,j-1} + 1 & \text{(Einfügung)} \\ M_{i-2,j-2} + 1, & \text{falls } S_{i-1} = Z_j \text{und } S_i = Z_{j-1} \text{ (Vertauschung)} \end{cases} . \tag{4.3}$$

Falls eine der aufgeführten Operationen nicht gebildet werden kann, da eine Zelle nicht definiert ist, wird die Operation nicht berücksichtigt. Z. B. ist eine Vertauschung für die Zelle $M_{1,1}$ nicht möglich.

Abbildung 4.11 zeigt ein Beispiel für eine Distanzmatrix mit dem Startwort DOROTHEA und dem Zielwort DOOHTER. Die Distanz beträgt 3. Das Traversieren der Matrix erfolgt durch eine Rekonstruktion der gewählten Operation. Eine vertikale Bewegung ohne Veränderung des Werts bedeutet, dass keine Operation durchgeführt wurde; eine vertikale Bewegung mit Veränderung des Werts bedeutet eine Ersetzung. Sofern die vorherigen beiden Buchstaben vertauscht werden können, bedeutet eine vertikale Bewegung mit Überspringen einer Zelle eine Vertauschung. Eine horizontale Bewegung bedeutet eine Einfügung und eine vertikale Bewegung eine Löschung. Die im Beispiel erfolgten Operationen sind eine Ersetzung (A zu R), eine Vertauschung (TH zu HT) und eine Löschung (R).

Länge	2	3–4	5–6	7–9	≥ 10
Kosten	1	2	3	4	5

Tabelle 4.6.: Kosten für das Auslassen eines Tokens nach Länge.

In manchen Fällen kann die Matrix nicht eindeutig traversiert werden. So können mehrere Kombinationen von Operationen die gleiche Distanz ergeben. Dieses Verhalten kann besonders bei vielen Fehler auftreten. Angesichts dessen ist die Reihenfolge der Abfrage relevant. Die implementierte Reihenfolge ist:

1. Keine Operation (Zeichen sind gleich),
2. Vertauschung,
3. Löschung,
4. Einfügung und
5. Ersetzung.

Die Ergebnisse des oben dargestellten Algorithmus sind nur gültig, falls zwei Schreibweisen des gleichen Tokens vorliegen. Daher erfolgt vor der Analyse der Tippfehler ein Abgleich der Token. Dies ist besonders bei Namen wichtig, da zwei Schreibweisen eines Namens aus einer unterschiedlichen Zusammensetzung von Token bestehen können.

Der Abgleich erfolgt, indem die Damerau-Levenshtein-Distanz für alle Permutationen von Token des Zielnamens und alle Kombinationen von Token des Startnamens berechnet wird. Hierzu werden die ausgewählten Token in der vorbestimmten Reihenfolge mit Leerzeichen aneinandergefügt. Die Token-Zusammensetzung mit der niedrigsten Distanz, den meisten Token und den wenigsten Vertauschungen in der Reihenfolge wird als Ausgangsbasis für den Vergleich verwendet. Das Auslassen eines Tokens wird abhängig von der Länge des Tokens mit Kosten bestraft (siehe Tabelle 4.6). Die Kosten geben die maximale Anzahl Fehler an, die in einem Token enthalten sein dürfen. Sie wurden durch Versuch und Irrtum ermittelt. Token mit nur einem Zeichen werden entfernt, da sie in den meisten Fällen eine Abkürzung und kein Tippfehler darstellen. Die Methode erkennt zudem auch Fehler, die auf unterschiedliche Trennungen zurückzuführen sind, da immer die Distanz der zusammengesetzten Zeichenketten berechnet wird.

4.5.2. Fehler in der Anzahl und Art verwendeter Token

Ein großer Anteil an Fehlern in Namen ist nicht auf Tippfehler, sondern auf eine unterschiedliche Menge an verwendeten Token zurückzuführen. 80.4 % der Unterschiede bei Vornamen enthalten keine Tippfehler, sondern nur unterschiedliche Mengen an Token. Bei Nachnamen sind dies 73 % und bei Blocknamen 72.8 %. Weitere 1.5 % der Vornamen sowie 2.6 % der Nachnamen und 3.2 % der Blocknamen enthalten sowohl Tippfehler als auch unterschiedliche Mengen an Token.[1] Insgesamt unterscheiden sich 3,035 Records im angegebenen Namen. 1,920 im Vornamen, 1,289 im Nachnamen und 2,850 im Blocknamen.

Es ist möglich, dass diese Fehler über die Art der Namen zu begründen sind. Zunächst ist es möglich, dass Namen mit vielen Token häufiger abgekürzt werden als Namen mit wenigen Token. Dies kann daran liegen, dass eine Angabe der weiteren Namen bei der Datenaufnahme als nicht notwendig erachtet wird. Der Zusammenhang wird in Tabelle 4.7 dargestellt. In allen Namensfeldern zeigt sich,

[1] Betrachtet wurden nur Records mit mindestens einem angegebenen Namen beim Vor-, Nach- oder Blocknamen.

	Vorname		Nachname		Blockname	
Anzahl Token	unvollständig (in %)	gesamt (in %)	unvollständig (in %)	gesamt (in %)	unvollständig (in %)	gesamt (in %)
0	0.000	0.264	0.000	0.008		0.000
1	0.001	50.254	0.001	94.230	0.000	0.015
2	0.003	42.218	0.022	4.887	0.002	47.358
3	0.025	6.487	0.267	0.617	0.003	42.741
4	0.332	0.661	1.220	0.135	0.026	8.370
5	2.210	0.099	2.247	0.098	0.136	1.212
6	0.000	0.014	9.756	0.022	0.895	0.245
7	0.000	0.002	16.667	0.003	2.299	0.048
8		0.000		0.000	0.000	0.009
9		0.000		0.000	0.000	0.002

Tabelle 4.7.: Verteilung der Vor-, Nach- und Blocknamen nach Anzahl Token und der Wahrscheinlichkeit für das Fehlen mindestens eines Tokens. Datenbasis für die Gesamtverteilung ist das ZSR 2023/24.

dass eine größere Anzahl an Token auch mit einer höheren Wahrscheinlichkeit für das Fehlen von Token einhergeht. So ist etwa die Wahrscheinlichkeit für das Fehlen eines Tokens bei Personen mit vier Vornamen mehr als 100-mal größer als bei einer Person mit zwei Vornamen.

Eine weitere Möglichkeit für unterschiedliche Token ist eine Änderung des Namens. Die Möglichkeiten zur Änderung des Nachnamens wurden bereits vorgestellt. Auch eine Änderung des Vornamens ist möglich, wenn auch rechtlich deutlich schwieriger. Neben einer tatsächlichen Änderung des Vornamens kann eine Veränderung insbesondere mit der Herkunft der Person zusammenhängen. Um eine Veränderung des Namens nach Herkunft quantifizieren zu können, lagen jedoch zu wenige Beobachtungen vor. Aus diesem Grund können nur einige qualitative Beschreibungen hierzu gegeben werden.

Eine auffällige Personengruppe mit Veränderungen im Vornamen sind Menschen aus dem Mittleren und Nahen Osten, insbesondere Syrien, Irak und Afghanistan. Diese Vornamen zeichnen sich besonders dadurch aus, dass Vor- und Nachnamen vertauscht wurden oder zwei sehr unterschiedliche Schreibweisen des gleichen Namens vorlagen. Eine solche Veränderung wird vom verwendeten Algorithmus nicht festgestellt, sodass die Token als unterschiedliche Namen behandelt werden.

Eine weitere auffällige Gruppe sind Menschen mit ostasiatischer Herkunft oder Abstammung. In diesen Ländern ist es oft üblich, Kindern zwei Namen zu geben: einen europäisierten Namen und einen Namen in der jeweiligen Landessprache. Unterschiede kommen dann zustande, wenn sich diese beiden Namen in den Datenbeständen abwechseln.

Beide genannten Gruppen stammen zudem aus Ländern mit einem anderen Schriftsystem. Entsprechend sind einige Fehler auch auf andere Transliterationen zurückzuführen.

Die Anzahl verwendeter Token unterscheidet sich auch nach Herkunft. Tabelle 4.8 stellt die Anzahl Token für Vor- und Nachname nach Herkunft dar (Schuljahr 2023/24). Es zeigt sich, dass Deutsche häufiger mehrere Vornamen besitzen. So haben 58.5 % der Deutschen mehrere Vornamen, jedoch nur 24.6 % der Migranten und 42.1 % der Menschen mit Migrationshintergrund. Etwa 10.9 % der Migranten und 10.8 % der Menschen mit Migrationshintergrund haben mehrere Nachnamen. Nur 2.9 % der Deutschen haben hingegen mehrere Nachnamen.

Für die Verknüpfung und auch für ein Clerical-Review sind die Namen wichtige distinguierende Merkmale. Hat eine Person mehrere Namen, so kann sie besser unterschieden werden. Hierbei ist

	Vorname (in %)			Nachname (in %)		
Anzahl Token	Mig.	Mig.-h.	Deut.	Mig.	Mig.-h.	Deut.
1	75.387	57.911	41.471	89.105	89.189	97.105
2	21.227	36.374	49.747	9.664	9.742	2.535
3	3.029	5.028	7.877	0.969	0.916	0.299
4	0.322	0.574	0.771	0.220	0.120	0.052
5	0.035	0.101	0.114	0.038	0.026	0.008
6		0.012	0.018	0.003	0.007	0.001
7			0.003			

Tabelle 4.8.: Anteil Anzahl Token in Vor- und Nachname nach Herkunft (Schuljahr 2023/24). Es wurden nur Melderegisterdaten berücksichtigt. Personen mit Blocknamen wurden ausgeschlossen.

auffällig, dass 67.2 % der Migranten nur einen Vor- und einen Nachnamen besitzen. Der Anteil beläuft sich hingegen bei Menschen mit Migrationshintergrund auf 52.9 % und bei Deutschen auf 45.2 %. Dies bedeutet, dass eine Unterscheidung von Migranten aufgrund des Namens schwieriger sein kann, insbesondere falls Fehler vorliegen. Aus diesem Grund wird im Folgenden die Namensvergabe nach Herkunft genauer betrachtet.

Zunächst wird die Gesamtverteilung der Namen betrachtet. Die normalisierte Entropie der Vornamen für das Schuljahr 2023/24 beträgt 0.91 bei Migranten, 0.93 bei Menschen mit Migrationshintergrund und 0.89 bei Deutschen. In dieser Reihenfolge beträgt die normalisierte Entropie der Nachnamen 0.95, 0.95 und 0.92 sowie die normalisierte Entropie der Blocknamen 0.9995, 0.9998 und 0.9997. Es muss beachtet werden, dass sich die Entropie durch Tippfehler erhöht. Daher muss auch der Anteil eindeutiger Namen betrachtet werden. 99 % der Blocknamen von Migranten sind eindeutig, 99.6 % von Menschen mit Migrationshintergrund und 99.2 % von Deutschen (Vornamen: 48.7 %, 57.4 %, 47.9 %; Nachnamen: 52 %, 49.2 %, 31.4 %). Der Anteil nicht eindeutiger Namen von Migranten ist dabei auffällig, denn aufgrund der geringen Anzahl Beobachtungen von Migranten ist zu erwarten, dass gleiche Namen unwahrscheinlicher sind.

Es muss zusätzlich beachtet werden, dass Migranten und Menschen mit Migrationshintergrund keine homogenen Gruppen sind, sondern verschiedene Ethnien umfassen. Dabei ist es möglich, dass es Ethnien mit einer sehr homogenen Namensvergabe gibt. Falls diese Ethnien geografisch eng beieinander leben (klumpen), kann dies zu einer schwierigeren Differenzierung der betroffenen Fälle führen. Daher wird der Anteil an Namen betrachtet, die innerhalb einer 100m-Gitterzelle nicht eindeutig sind. Bei Blocknamen beträgt der Anteil für Migranten 0.015 %, für Menschen mit Migrationshintergrund 0.002 % und für Deutsche 0.001 %. Bei Vornamen beträgt der Anteil für Migranten 2.8 %, für Menschen mit Migrationshintergrund 0.5 % und für Deutsche 0.2 %. Konkret bedeutet die Werte, dass z. B. für einen zufällig ausgewählten Migranten mit einer Wahrscheinlichkeit von 2.8 % innerhalb derselben 100m-Gitterzelle ein weiterer Migrant im ZSR enthalten ist, der den gleichen Vornamen hat. Da der Anteil gleicher Nachnamen innerhalb einer 100m-Gitterzelle maßgeblich durch die Anzahl Geschwister beeinflusst wird, sind diese Anteile deutlich höher als bei Vornamen. Der Anteil beträgt für Migranten 26.4 %, für Menschen mit Migrationshintergrund 23.9 % und für Deutsche 25.8 %. Alle Ergebnisse zeigen, dass Migranten mit gleichen Namen häufiger geografisch klumpen als es bei Deutschen oder Menschen mit Migrationshintergrund der Fall ist. Hieraus folgt eine höhere Verwechslungswahrscheinlichkeit, die weitere Probleme bei der Verknüpfung von Migranten hervorrufen kann.

Für die Namensvergabe wurde zusätzlich ein Kohorteneffekt untersucht. Abbildung 4.12 zeigt den Anteil der Anzahl Token im Vornamen nach Herkunft und Geburtsjahr. Ein starker Umbruch kann

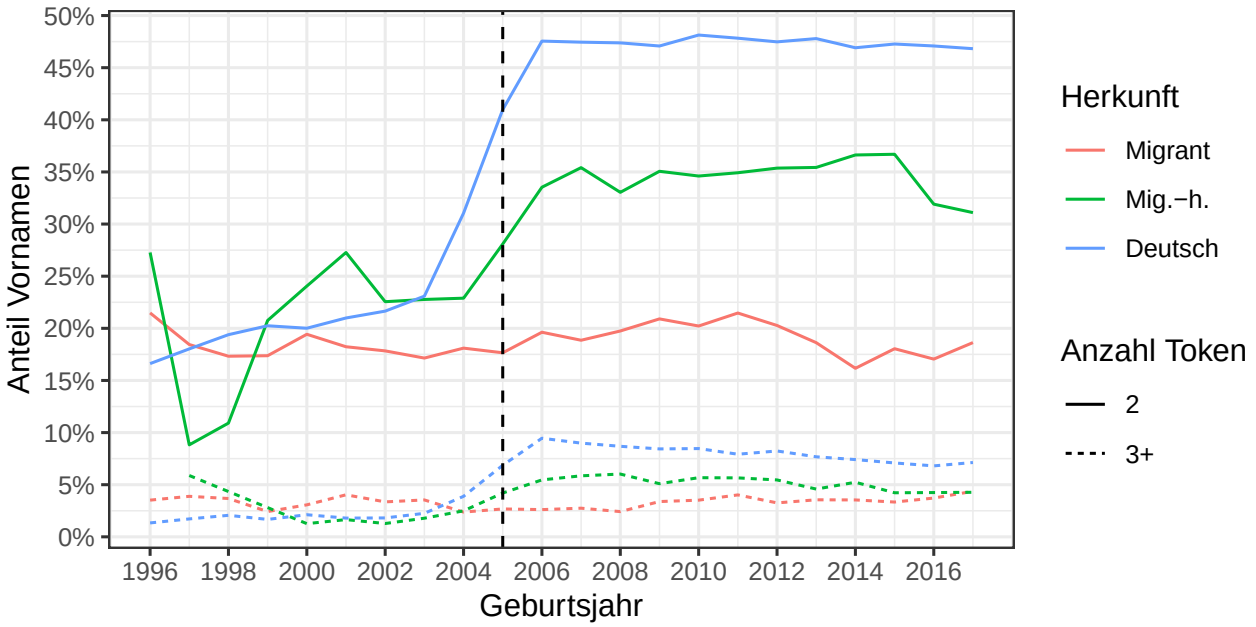


Abbildung 4.12.: Anteil Vornamen nach Geburtsjahr, Herkunft und Anzahl Token (Schuljahr 2023/24). Vornamen mit drei oder mehr Token sind zusammengefasst. Vornamen mit einem Token sind nicht dargestellt (die Werte ergeben sich aus den fehlenden Anteilen). Die vertikale, gestrichelte Line kennzeichnet das Jahr 2005. Beobachtungen, die in diesem Jahr oder früher geboren wurden, sind mindestens 18 Jahre alt und wurden somit von der Schule erfasst (Beobachtungen links von der Linie).

bei Personen, die mindestens 18 Jahre als sind, beobachtet werden. Dies zeigt, dass bei der Erfassung der Namen in den Schulen meist nur der Rufname erfasst wird. Für Personen, die jünger als 18 Jahre sind, ist der Anteil Token relativ konstant. Ein leichter Kohorteneffekt ist bei der Vergabe von mehr als zwei Token bei Deutschen zu beobachten. Der Anteil sank innerhalb von 10 Jahren linear um insgesamt 3 %.

Bei der Betrachtung des Anteils Vornamen mit Bindestrich zeigt sich ein weiterer Kohorteneffekt (Abbildung 4.13). So sinkt der Anteil bei Deutschen seit 1998 nahezu linear um 3 %. Im Zusammenhang zeigt sich, dass Vornamen mit mehreren Token immer häufiger getrennt als mit Bindestrich vergeben werden.

Auf eine Darstellung der Anteile Token im Nachnamen wird verzichtet. Weder ein Kohorteneffekt noch ein Alterseffekt ist dabei erkennbar. Es zeigt sich jedoch, dass der höhere Anteil Nachnamen mit zwei Token bei Migranten und Menschen mit Migrationshintergrund zu einem großen Teil auf herkunftsspezifische Präfixe zurückzuführen ist (z. B. Al, Ben, De, Dos, El, La, Van). 5.7 % aller Nachnamen von Migranten und Menschen mit Migrationshintergrund mit mehr als einem Token enthalten z. B. das Präfix “Al”. Dennoch besteht die Möglichkeit, dass diese Gruppe häufiger Doppelnamen annimmt. Eine Analyse ist im Rahmen dieser Arbeit jedoch nicht möglich.

Die Ergebnisse zeigen, dass die Veränderung in der Tokenzahl, insbesondere aufgrund der Meldung durch die Schule, weiter betrachtet werden muss. Abbildungen 4.14, 4.15 und 4.16 zeigen daher die relative Veränderung der Anzahl übereinstimmender Token, falls sich die Namen zwischen den Datensätzen unterscheiden. Sowohl Vor- als auch Nachnamen stimmen unabhängig von der Anzahl Token meist nur in einem Token überein. Dies bedeutet, dass häufig eine Reduktion von vielen Token auf ein

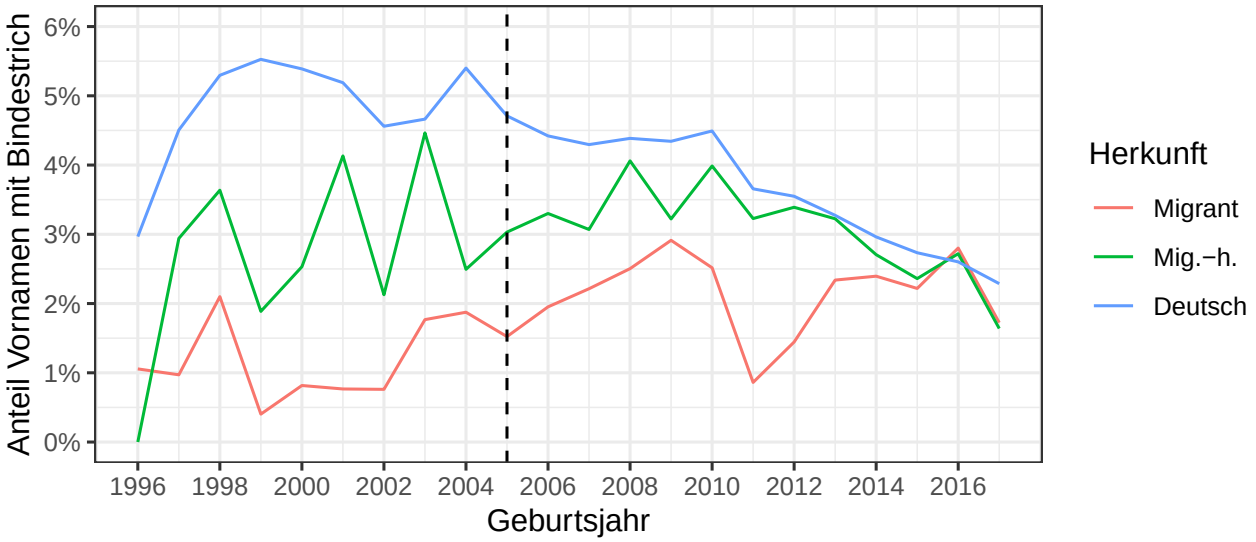


Abbildung 4.13.: Anteil Vornamen mit Bindestrich nach Geburtsjahr und Herkunft (Schuljahr 2023/24). Die vertikale, gestrichelte Line kennzeichnet das Jahr 2005. Beobachtungen, die in diesem Jahr oder früher geboren wurden, sind mindestens 18 Jahre alt und wurden somit von der Schule erfasst.

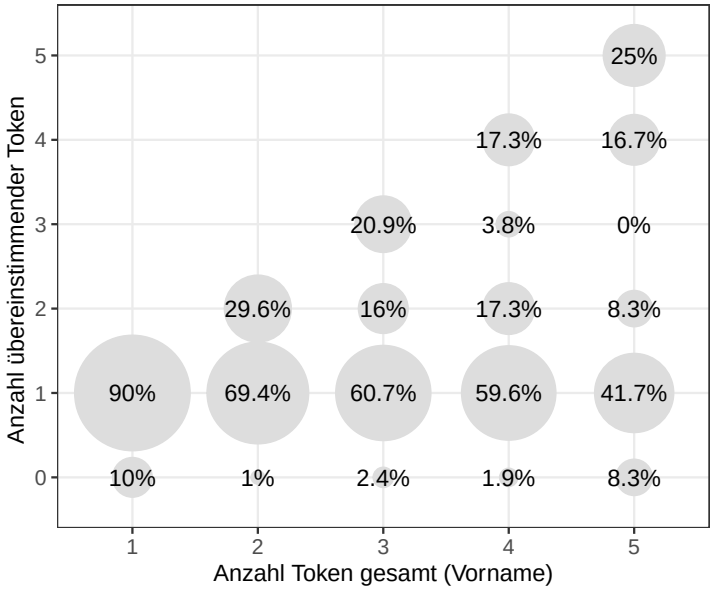


Abbildung 4.14.: Verhältnis der Gesamtanzahl Token zur Anzahl übereinstimmender Token bei Vornamen. Betrachtet werden nur Vornamen, die sich zwischen den Schuljahren unterscheiden.

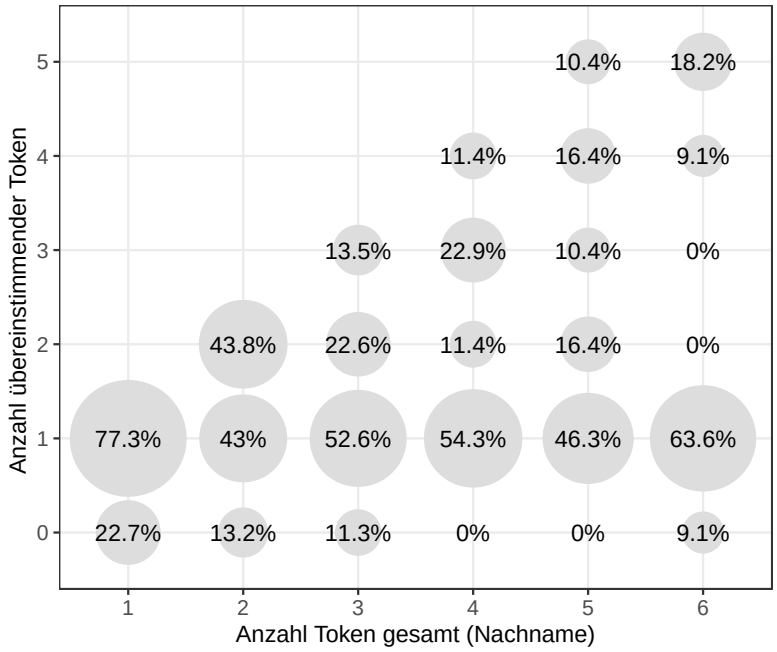


Abbildung 4.15.: Verhältnis der Gesamtanzahl Token zur Anzahl übereinstimmender Token bei Nachnamen. Betrachtet werden nur Nachnamen, die sich zwischen den Schuljahren unterscheiden. Nachnamen mit mehr als sechs Token werden aufgrund einer zu geringen Fallzahl nicht dargestellt.

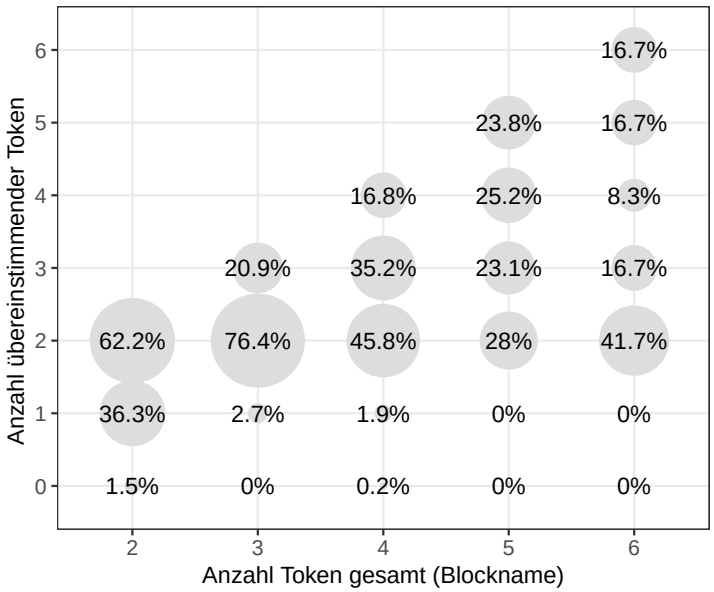


Abbildung 4.16.: Verhältnis der Gesamtanzahl Token zur Anzahl übereinstimmender Token bei Blocknamen. Betrachtet werden nur Blocknamen, die sich zwischen den Schuljahren unterscheiden. Blocknamen mit mehr als sechs Token werden aufgrund einer zu geringen Fallzahl nicht dargestellt.

Token stattfindet (z. B. von PETER UWE zu PETER). Die Hauptdiagonale zeigt jeweils den Anteil Namen, die ausschließlich Tippfehler beinhalten (z. B. PETE UWE und PETER UWE).

Falls kein Token des Nachnamens übereinstimmt, so ist dies ein Zeichen für einen möglichen Wechsel des Nachnamens. Von allen festgestellten Nachnamen stimmen insgesamt 20.2 % in keinem Token überein. Ein Namenswechsel ist damit wahrscheinlich (siehe zur Gesamtverteilung Tabelle 4.7). Bei Vornamen beträgt der Anteil hingegen 3.7 %.

Die Verteilung der Anzahl übereinstimmender Token besitzt bei Blocknamen eine höhere Varianz als bei einer einzelnen Betrachtung von Vor- und Nachnamen. Dies kann damit begründet werden, dass sowohl die Anzahl Vornamen als auch die Anzahl Nachnamen variieren kann. Dennoch zeigt sich, dass unabhängig von der Anzahl der Token meist nur zwei Token übereinstimmen (ein Vor- und ein Nachname). Die Wahrscheinlichkeit, dass kein Token übereinstimmt, ist mit insgesamt 0.3 % sehr gering. Dennoch ist dies der schlecht möglichste Fall bei einer Verknüpfung. 8 der 10 Fälle waren Migranten, 2 Fälle waren Menschen mit Migrationshintergrund. Es zeigt sich bei Betrachtung der Namen eine weit entfernte Ähnlichkeit zwischen den Namen in beiden Schuljahren.

4.5.3. Tippfehleranalyse

Im Folgenden wird die Verteilung von Tippfehlern für Vor-, Nach- und Blocknamen sowie Geburtstag, -monat, -jahr und Postleitzahl untersucht. Beim Vergleich der Namen und des Geburtstags liegen jeweils nur zwei Werte vor. Es wurde angenommen, dass der Datensatz für das Schuljahr 2021/22 eher den korrekten Namen enthält, da viele Fehler bei der Veränderung der Datenquelle vom Melderegister zur Schule entstehen. In diesem Fall liegen im Schuljahr 2021/22 die Angaben aus dem Melderegister vor und im Schuljahr 2023/24 die Angaben von der Schule. Für den Geburtsmonat und das Geburtsjahr liegen neben den Angaben aus dem ZSR auch Angaben aus dem IVDS vor. In diesem Fall wurde die am häufigsten genannte Zahl als korrekt angesehen. Falls mehrere Zahlen gleich häufig vorkommen, wurde die Angabe aus dem Schuljahr 2021/22 als wahrer Wert verwendet. Migranten, die am 1. 1. geboren wurden, wurden bei der Analyse von Fehlern im Geburtsdatum ausgeschlossen. Wie im Abschnitt 4.4.3 beschrieben, handelt es sich hierbei wahrscheinlich nicht um Tippfehler, sondern um eine unvollständige Angabe.

Zur Identifikation von Tippfehlern in Postleitzahlen wurden zwei Varianten verwendet. Als Erstes wurde die Veränderung der Postleitzahl zwischen beiden Schuljahren betrachtet. Da ein Umzug auch eine Veränderung der Postleitzahl bewirkt, wurden nur Veränderungen berücksichtigt, bei der sich Straße und Hausnummer nicht ändern.[1] Die zweite Variante bestand darin, nicht vergebene Postleitzahlen zu identifizieren und den Wert mit der richtigen Postleitzahl zu vergleichen. Die richtige Postleitzahl wurde anhand der zugewiesenen Geo-Koordinate und den entsprechenden geografischen Flächen der Postleitzahlgebiete gewonnen (siehe Abschnitt 4.2).

Das beschriebene Vorgehen wurde auch mit den gültigen Postleitzahlen getestet. Die Ergebnisse zeigten jedoch, dass die Datenqualität von OSM nicht ausreichend ist. Eine Vielzahl der so festgestellten Fehler beruhten auf einer fehlenden Adresse in den OSM-Daten und nicht auf einer falschen Postleitzahlzuordnung. Da der Aufwand zur Identifizierung aller fehlerhaften Adressen zu hoch war, wurde von einer solchen Analyse abgesehen. Für eine Untersuchung der ungültigen Postleitzahlen werden die Zuordnungen als ausreichend tragfähig erachtet, da kein Zusammenhang zwischen der Angabe einer ungültigen Postleitzahl und dem Fehlen der Adresse in den OSM-Daten besteht.

[1] Eine Änderung der Postleitzahl aufgrund einer Veränderung des Postleitzahlgebietes könnte auch einen Unterschied bedingen. Bei keiner der identifizierten falschen Postleitzahlen erfolgte im Untersuchungszeitraum eine Veränderung der Gebietszuordnung (Deutsche Post 2024; OpenStreetMap Wiki 2023).

	Anteil mit f Fehlern (in %)							
Variable	$f=1$	$f=2$	$f=3$	$f=4$	$f=5$	$f=6$	ausg. (in %)	n
Vorname (gesamt)	91.0	6.9	1.3	0.8				377
Nachanme (gesamt)	83.9	13.2	1.1	1.1	0.3	0.3		348
Blockname (gesamt)	77.8	14.3	4.9	2.6	0.3	0.1		776
Vorname (Token)	90.9	7.6	1.3	0.3				384
Nachanme (Token)	85.1	13.2	1.1	0.6				363
Blockname (Token)	80.6	15.0	3.7	0.7				842
Geburtstag	72.7	27.3					27.3	187
Geburtsmonat	86.3	13.7					13.7	190
Geburtsmonat (+IVDS)	85.4	14.6					14.6	260
Geburtsjahr	83.6	4.8	1.4	10.3			16.4	146
Geburtsjahr (+IVDS)	85.8	6.0	1.1	7.1			14.2	183
PLZ (wechselnd)	69.1	15.5	12.4	3.1			15.5	97
PLZ (nicht vergeben)	70.2	14.9	14.0	0.8			9.8	121
PLZ (gesamt)	68.5	15.3	14.3	2.0			12.5	203

Tabelle 4.9.: Verteilung der Anzahl Fehler (Damerau-Levenshtein-Distanz) nach Variable. Die Zusammensetzung der identifizierten übereinstimmenden Token eines Namens bilden den gesamten Namen. Die Spalte ausg. stellt den Anteil Fälle dar, die durch die Begrenzung auf eine maximale Anzahl Fehler ausgeschlossen werden (maximal ein Fehler bei Geburtstag, -monat und -jahr; maximal zwei Fehler bei der Postleitzahl).

Um weitere Fehler auszuschließen, die nicht auf Tippfehler zurückzuführen sind, wurden, ähnlich wie bei den Namen, Begrenzungen für eine maximale Anzahl Fehler festgelegt. Die maximale Damerau-Levenshtein-Distanz für Geburtstag, -monat und -jahr wurde auf eins festgelegt, für die Postleitzahl auf zwei.

Tabelle 4.9 stellt die Anteile der Anzahl Fehler dar sowie den Anteil der Fälle, die durch das Einführen der Maximaldistanz für eine weitere Analyse ausgeschlossen werden. Dabei werden mehrere Unterscheidungen eingeführt. Namen werden zwischen dem gesamten zusammengesetzten Namen und den einzelnen Token unterschieden. Es muss beachtet werden, dass die Anzahl Fehler im Gesamtnamen nicht zwangsläufig die Summe der Fehler der einzelnen Token ist. Dies ist z. B. bei einer Vertauschung eines Leerzeichens der Fall. Ferner wird für den Geburtsmonat und das Geburtsjahr zwischen den Datenquellen unterschieden. Zum einen werden nur die Fehler im ZSR betrachtet und zum anderen die Gesamtmenge an Fehlern im ZSR und IVDS. Die Postleitzahl wird nach den beiden Varianten sowie der Gesamtmenge aus beiden Varianten unterschieden.

Zwischen den verschiedenen Mengendefinitionen einer Variablen können nur geringe Unterschiede in der Fehlerverteilung festgestellt werden. Die mit Abstand meisten Token haben einen Fehler. Insbesondere die hohen Zahlen für einen vollständigen Wechsel des Geburtstags (27.3 %) und des Geburtsjahrs (10.3 %) im ZSR fallen jedoch auf. Bei einer genaueren Untersuchung dieser Fälle kann kein Zusammenhang zur Herkunft oder Datenquelle festgestellt werden. Auffällig bei Geburtstagen ist, dass in 25.5 % dieser Fälle der Eintrag im Schuljahr 2021/22 der 1. ist. Ein derartiger Zusammenhang kann bei den Geburtstagen für das Schuljahr 2023/24 nicht beobachtet werden. Auffällig ist nur ein hoher Anteil von Records, die durch das Clerical-Review zusammengeführt wurden (23–47%). Dies spricht nicht gegen die Ergebnisse des Clerical-Reviews, sondern zeigt nur, dass falsche Geburtsdaten bei der Verknüpfung des ZSR häufig zu einem false negative führen.

	Fehlerquote (in %)				Fehlerarten (in %)				
Variable	Gesamt	Meld.	Schule	Beide	Ersetzt	Gelö.	Eingef.	Vert.	n
Vorname (gesamt)	0.21	0.08	1.06	0.70	37.9	32.2	25.6	4.3	377
Nachname (gesamt)	0.19	0.10	0.73	0.58	34.0	34.8	27.4	3.8	348
Blockname (gesamt)	0.43	0.20	1.83	1.37	38.3	32.2	26.2	3.3	776
Geburtstag	0.10	0.05	0.50	0.26	76.5	8.1	14.7	0.7	136
Geburtsmonat	0.10	0.04	0.62	0.29	82.3	8.5	9.1	0.0	164
Geburtsmonat (+IVDS)					86.0	7.7	6.3	0.0	222
Geburtsjahr	0.08	0.02	0.55	0.24	99.2	0.0	0.0	0.8	122
Geburtsjahr (+IVDS)					99.4	0.0	0.0	0.6	157
PLZ (wechselnd)	0.05	0.00	0.38	0.28	70.1	6.2	6.2	17.5	82
PLZ (nicht vergeben)					62.8	8.3	8.3	20.7	103
PLZ (gesamt)					66.2	7.5	7.5	18.9	170

Tabelle 4.10.: Verteilung der Tippfehler und Fehlerquoten. Die Fehlerquoten berechnen sich aus der Anzahl Veränderungen, dividiert durch die Anzahl Records in beiden Datensätzen. Bei nicht angegebenen Fehlerquoten handelt es sich um zusammengesetzte Merkmale.

Tabelle 4.10 stellt die Verteilung der Tippfehlerarten und Fehlerquoten dar. Fehlerquoten werden dabei nur für den Vergleich zweier Merkmale des ZSR angegeben, da die anderen Vergleiche aus einer gemischten Verteilung stammen.[1] Es zeigt sich deutlich, dass sich die Verteilung der Fehlerarten für Namen von Fehlern in Zahlenfeldern unterscheiden. So weisen alle Zahlenfelder zu einem großen Teil Ersetzungen auf und nur sehr wenige Löschungen und Einfügungen. Namen weisen hingegen deutlich mehr Löschungen und Ersetzungen auf, wobei auch hier die meisten Fehler auf Ersetzungen zurückzuführen sind. Vertauschungen sind bei Namen mit etwa 3–4 % selten und bei Geburtsdaten mit etwa 0–1 % sehr selten. Dagegen stellen Vertauschungen bei Postleitzahlen mit ungefähr 19 % die zweithäufigste Fehlerursache dar. Hierbei muss beachtet werden, dass eine Vertauschung bei Geburtstag, -monat oder -jahr häufig zu ungültigen Werten führt. So existiert für den Monat keine Vertauschung, die zu einem gültigen Wert führen würde.[2] Dies ist bei Postleitzahlen nicht der Fall. Die Verteilung der Fehlerarten für Geburtstag, -monat oder -jahr gehen somit auch darauf zurück, dass häufig kein anderer Fehler als eine Ersetzung möglich ist. Dies zeigt, dass den einzelnen Variablen unterschiedliche Fehlercharakteristiken zugrunde liegen.

Die Verteilung der Fehlerarten lässt sich auch mit einigen in der Literatur beschriebenen Verteilungen vergleichen. Eine der größten Studien zu Fehlern in Wörtern wurde von Pollock & Zamora (1983; 1984) durchgeführt. Datengrundlage waren 25 Mio. Wörtern aus sieben wissenschaftlichen Zeitschriften, welche mit einem Wörterbuch mit 40,000 Wörtern abgeglichen wurden. Dabei wurden 19 % Ersetzungen, 34 % Löschungen, 27 % Einfügungen und 12.5 % Vertauschungen festgestellt. 7.5 % der Wörter wiesen mehrere Fehler auf, welche die Autoren nicht klassifizierten.

[1] Aufgrund der unterschiedlichen Grundgesamtheit hätte die Angabe der Fehlerquoten die Komplexität der Ergebnisdarstellung zu sehr erhöht. Die Fehlerquoten sind jedoch ähnlich wie die Fehlerquoten, die auf Grundlage des ZSR beschriebenen wurden.

[2] Theoretisch ist eine Vertauschung des Monats Januar (01) und Oktober (10) möglich. In der Analyse werden jedoch keine führenden Nullen angegeben. Eine Verwendung von führenden Nullen würde zudem zu zahlreichen Problemen bei anderen Fehlern führen würde. So würde z. B. eine Löschung vom Monat 12 auf 01 zwei Fehler verursachen (Löschung der Zahl 2 und Einfügung der Zahl 0). Aus diesem Grund sind für die betrachteten Daten keine Vertauschungen des Monats möglich.

Nur sehr wenige Studien befassen sich mit Fehlern in Nicht-Wörtern, wie Namen von Personen oder Orten (Tagliacozzo et al. 1970; Lenman & Marmolin 1987). Eine geeignete Zahl liefern ausschließlich Tagliacozzo et al. (1970). Bei einer Befragung von 1489 Bibliotheksbesuchern wurden insgesamt 104 Fehler in 87 Autorennamen identifiziert, welche von den Besuchern nicht im Bibliothekskatalog gefunden werden konnten. Die Fehlerarten verteilten sich auf 48.1 % Ersetzungen, 32.7 % Löschungen, 17.3 % Einfügungen und 2 % Vertauschungen.

Die größte Datenbasis für Tippfehler in Namen bietet eine noch unveröffentlichte Studie von Schnell & Weiand. Beim Vergleich zwischen zwei Literaturdatenbanken mit insgesamt über 1,2 Mio. gemeinsamen Einträgen konnten 48,553 Tippfehler identifiziert werden. Die Fehler verteilen sich auf 38.4 % Ersetzungen, 36.6 % Löschungen, 23.6 % Einfügungen und 1.3 % Vertauschungen.

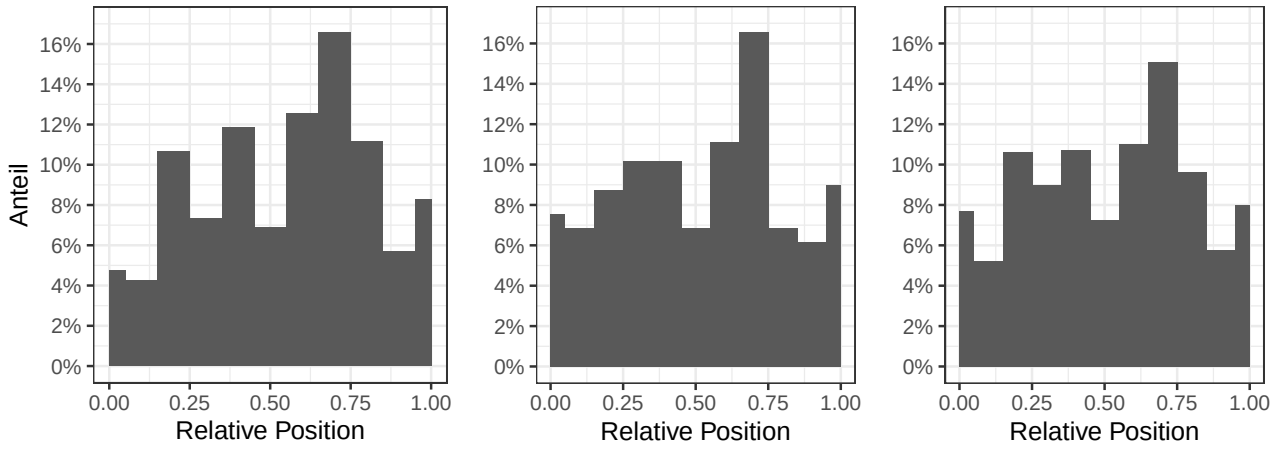
Für die Simulation von Tippfehlern bei der Mikrosimulation des Bildungsverlaufsregisters wurden die Ergebnisse von Peterson (1986) verwendet. Auf Basis von 369,546 abgetippter Wörter aus 16 Dokumenten wurden 43.1 % Ersetzungen, 34 % Löschungen, 20.1 % Einfügungen und 2.8 % Vertauschungen festgestellt. Diese Zahlen liegen nahe an den empirischen Ergebnissen. Die Simulation der Tippfehlerarten stellte somit eine gute Approximation dar.

Es zeigt sich zum einen, dass die Verteilung der Fehlerarten ähnlich zu anderen Studien insbesondere mit Nicht-Wörtern sind. Eine große Ähnlichkeit besteht besonders zu den Ergebnissen von Schnell & Weiand, welche zudem auf einer großen Fallzahl beruhen. Zum anderen zeigt sich, dass die Ergebnisse von empirischen Ergebnissen zu Wörtern zum Teil stark abweichen. Es kann somit gefolgert werden, dass Fehler in Namen sich von Fehlern in Wörtern unterscheiden. Dies muss für Simulationen berücksichtigt werden. Auch unterscheiden sich diese Verteilungen wiederum zu Fehlern in Zahlenfeldern. Dabei gibt es keine Literatur zu Fehlern in Zahlen. Die hier präsentierten Ergebnisse sind somit eine Neuheit und können nicht mit anderen Ergebnissen verglichen werden.

Für die Qualität von phonetischen Codierungen, wie z. B. Soundex, sowie von Ähnlichkeitsmaßen, wie z. B. Jaro-Winkler, ist die Position der Fehler relevant. So gewichten sowohl Soundex als auch die Jaro-Winkler-Ähnlichkeit den Anfang einer Zeichenkette stärker (Christen 2012: 74, 109). Abbildungen 4.17a und 4.17b stellen die Verteilung der Fehlerposition relativ zur Länge des Vor- und Nachnamens für alle Fehlerarten dar. Es zeigt sich, dass die Fehler in Vor- und Nachnamen etwas anders verteilt sind. So sind die Fehler bei Nachnamen häufiger näher am Anfang als bei Vornamen. Der Anteil Fehler am ersten Buchstaben beträgt bei Nachnamen 7.6 % und 4.7 % bei Vornamen. Aufgrund der geringen Zahl Beobachtungen ist dieser Unterschied nicht signifikant.[1] Abbildung 4.17c zeigt daher die gemeinsame Verteilung der Fehlerposition für Vor- und Nachnamen durch Verwendung des Blocknamens. Auch hier zeigt sich, dass Fehler häufig in der Mitte, tendenziell aber weiter am Ende als am Anfang auftreten. Zudem fällt auf, dass der Anteil Fehler in der ersten Position bei 7.7 % liegt und damit sogar höher ist als bei Nachnamen. Ursache hierfür sind falsch getrennte Namen, wobei ein Teil dem Vornamen und der andere dem Nachnamen zugeordnet wurde.

Die dargestellten Verteilungen decken sich mit einigen in der Literatur genannten Verteilungen. So stellten auch (Pollock & Zamora 1983) fest, dass Fehler in Wörtern tendenziell weiter am Ende auftreten. Christen & Pudjijono (2009) stellten wiederum fest, dass 8 % aller Fehler den ersten Buchstaben betreffen. Diese Zahl wurde auch für die Simulation angenommen, wobei eine Gleichverteilung für die anderen Positionen angenommen wurde. Im Vergleich zu den empirischen Daten war dies somit eine gute Approximation.

[1] Getestet mit einem zweiseitigen Test für Anteilswerte (R-Funktion `prop.test`), bei $\alpha = 0.05$.



(a) Vorname (b) Nachname (c) Blockname

Abbildung 4.17.: Anteil Fehler nach relativer Position in Vor-, Nach- und Blocknamen für alle Fehlerarten. Die relative Position p_r ist definiert als das Verhältnis der absoluten Position des Fehlers im Startwort p_S zur Länge des Startworts l_S: $p_r = p_S/l_S$. Fehler an der ersten und letzten Stelle werden in separaten Balken dargestellt.

Für die Untersuchung der Zahlenwerte ergibt sich der Vorteil, dass die Werte eine feste maximale Anzahl Stellen besitzen. Daher kann die absolute Position des Fehlers untersucht werden.[1] Die Fehlerposition wird für eine konsistente Darstellung anhand der maximalen Anzahl Ziffern eines Zahlenwertes bestimmt. Dies hat zur Folge, dass für den Geburtstag und Monat in einigen Fällen kein Wert für die erste Stelle vergeben ist (z. B. der Monat März hat nur die Zahl 3 an der zweiten Stelle). Ein Problem bei einer solchen Form der Untersuchung stellen Vertauschungen dar, da diese immer zwei Stellen betreffen. Vertauschungen werden daher separat behandelt.

Aufgrund der beschriebenen Einschränkungen ist die Position der Fehler bei Geburtsdaten stark determiniert. Dies hat zur Folge, dass 97 % der Fehler (ohne Vertauschungen) beim Geburtsmonat und 85 % beim Geburtstag an der zweiten Position liegen. Auch das Geburtsjahr kann nur in den letzten zwei Stellen variieren.[2] So liegen 98 % der Fehler an der letzten (vierten) Stellen. Die wenigen Vertauschungen sind immer auf die letzten zwei Stellen beschränkt.

Einzig Postleitzahlen können an jeder Position frei variieren. Abbildung 4.18 zeigt die Anteile Fehler nach Stelle (ohne Vertauschungen). Die Fehler konzentrierten sich besonders auf die letzten beiden Stellen. Vertauschungen verteilen sich etwa gleich häufig auf Vertauschungen der vierten und fünften Stelle (52.6 %) und der dritten und vierten Stelle (47.4 %). Bei allen Zahlenfeldern zeigt sich somit insgesamt eine Konzentration der Fehler auf die letzten Stellen.

[1] Für die Untersuchung der Position wurden jeweils die gemeinsamen Ergebnisse aus ZSR und IVDS für den Geburtsmonat und das Geburtsjahr verwendet. Für die Postleitzahl wurden die gemeinsamen Ergebnisse aus wechselnden und ungültigen Postleitzahlen verwendet.

[2] Da nur syntaktisch genaue Geburtsjahre im Datensatz enthalten sind, muss für einen Fehler am Anfang der Zahl immer die ersten beiden Stellen verändert werden (von 19 zu 20 oder andersherum). Eine solche Veränderung würde somit zwei Fehler verursachen. Gemäß der aufgestellten Begrenzungen, dass beim Geburtsjahr maximal ein Fehler vorliegen darf, werden diese Fälle somit ausgeschlossen. Dadurch sind beim Geburtsjahr und Fehler an den letzten zwei Stellen möglich.

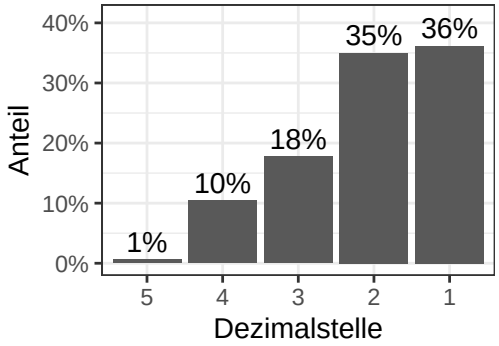


Abbildung 4.18.: Anteil Fehler (ohne Vertauschungen) der Postleitzahl (gesamt) nach Position an der die Fehler auftreten.

Als Letztes können Tippfehler auch dahin gehend untersucht werden, welche Fehlerart häufig bei bestimmten Zeichen auftreten. Gegeben der vielen möglichen Kombinationen von Buchstaben und der Verteilung der Fehlerarten liegen für eine derartige Untersuchung bei Namen zu wenig Beobachtungen vor. Eine Untersuchung für Ersetzungen bei Zahlenwerten ist hingegen möglich, da die meisten entdeckten Fehler bei Zahlen Vertauschungen sind und es weniger Kombinationen gibt.

Aufgrund der wenigen Fehler in Zahlenfeldern mussten einige Annahmen zur Untersuchung der Ersetzungen getroffen werden. Zunächst wurde angenommen, dass die bidirektionalen Ersetzungswahrscheinlichkeiten gleich sind. Dies bedeutet, dass die Ersetzungswahrscheinlichkeiten von x mit y genauso hoch ist wie die Ersetzungswahrscheinlichkeiten von y mit x. Hieraus reduziert sich die Anzahl möglicher Ersetzungen um die Hälfte. Als Nächstes wurde angenommen, dass die Ersetzungen für alle frei variierenden Dezimalstellen der gleichen Verteilung folgen. Somit können alle Zahlenfehler an der ersten Dezimalstelle des Geburtstags, -monats und -jahrs sowie die ersten zwei Stellen der Postleitzahl zusammengefasst werden. Daraus resultiert eine Mischverteilung, sodass theoretisch unterschiedliche Wahrscheinlichkeiten für Zahlenvertauschungen vorliegen. Das Problem der Darstellung einer Wahrscheinlichkeit auf Grundlage dieser Verteilung ist eine mögliche Überinterpretation der Ergebnisse, welche auf einer geringen Fallzahl beruhen. Stattdessen wurde daher eine Gleichverteilung der möglichen Ersetzungen angenommen. Dies erlaubt eine Interpretation der empirisch beobachteten Häufigkeiten.

Abbildung 4.19 stellt die bidirektionalen Häufigkeiten der Vertauschungen dar. Unter den genannten Annahme ergibt sich ein Erwartungswert von 5.7 Ersetzungen je Kombinationen. Alle Kombinationen, die über diesem Erwartungswert liegen, werden in der Abbildung farblich orange und rot gekennzeichnet. Die Abbildung zeigt, dass die Hauptdiagonale eine erhöhte Häufigkeit aufweist. Dies bedeutet, dass benachbarte Zahlen eine höhere Ersetzungswahrscheinlichkeit als weiter entfernte Zahlen haben. Eine Ursache hierfür kann eine physische Nähe der Tasten auf der Tastatur sein. Da Zahlen sowohl über einen Nummernblock als auch über die Zahlenreihe eingegeben werden können, lassen sich die Fehler nicht direkt auf eine Eingabeform zurückführen. Auf weitere Annahmen hierzu wird verzichtet. Die Ergebnisse zeigen jedoch, dass auch Ersetzungen in Zahlen nicht zufällig verteilt sind.

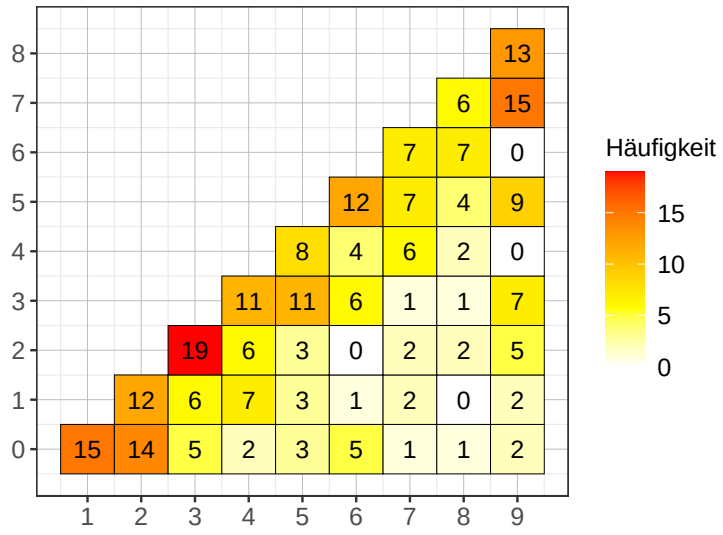


Abbildung 4.19.: Häufigkeit von Ersetzungen eines Zahlenpaars.

4.5.4. Geografische Fehler

Falls ein Fehler aufgrund der geografischen Nähe zweier Objekte zustande kommt, soll von einem geografischen Fehler gesprochen werden. Dieser ist insbesondere bei Postleitzahlen möglich und erkennbar. So besteht die Möglichkeit, dass bei der Eingabe einer Postleitzahl nicht die korrekte, sondern eine nahe gelegene angegeben wird. Gründe hierfür können vielfältig sein und hängen vom Eingabeprozess sowie der Validierung ab. Z. B. ist es möglich, dass die Postleitzahl anhand der Adresse oder einer Ortsangabe geraten wird.[1] In diesen Fällen kann erwartet werden, dass die Gebiete meist aneinandergrenzen.

Aus den genannten Gründen kann eine Analyse nur auf Grundlage der Postleitzahlen erfolgen, die wechseln, während Straße und Hausnummer unverändert bleiben. Eine Verteilung der Distanz zwischen den beiden Postleitzahlgebieten wird in Abbildung 4.20 dargestellt.[2] Die Distanz beschreibt die minimale Anzahl Gebietsgrenzen, die zu überschreiten sind, um von der ersten zur zweiten Postleitzahl zu gelangen. Eine Distanz von 1 besagt somit, dass die Gebiete direkt aneinander liegen. Die Abbildung zeigt, dass geografische Fehler durchaus wahrscheinlich sind. Bei 46 % der falsch angegebenen Postleitzahlen handelt es sich um benachbarte Gebiete. Obwohl die Ergebnisse auf einer geringen Fallzahl beruhen, zeigt dies dennoch, dass auch die Postleitzahl mit weiteren nicht zufälligen Fehlern behaftet ist.

[1] Auch ist es möglich, dass sich der Postleitzahlbereich geändert hat. Dies war jedoch für keinen der gefundenen Fehler der Fall.

[2] Technisch wurde das Problem über einen Graphen gelöst. Die Knoten stellen jeweils Postleitzahlgebiete dar, welche mit den Kanten zu allen benachbarten Gebieten verbunden sind. Die Informationen stammen aus den Polygonen der Postleitzahlgebiete (siehe Abschnitt 4.2). Die Entfernung ist damit die minimale Distanz zwischen den Konten im Graphen.

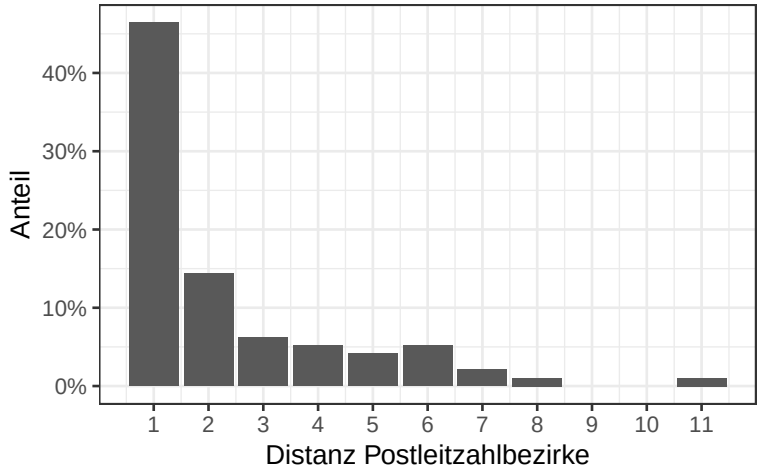


Abbildung 4.20.: Distanz zwischen zwei gültigen Postleitzahlen, bei unveränderter Straße und Hausnummer ($n = 83$). Die Distanz beschreibt die minimale Anzahl Gebietsgrenzen, die zu überschreiten sind, um von der ersten zur zweiten Postleitzahl zu gelangen.

4.6. Zwischenfazit

Die bisherigen Ergebnisse zeigen, dass Unterschiede in der Datenqualität nach Herkunft vorhanden sind. Insbesondere Migranten weisen höhere Fehlerquoten im Vergleich zu Deutschen auf. Hinzu kommt, dass sich ein unterschiedliches Umzugsverhalten nach Herkunft zeigt. So ziehen besonders Migranten häufiger und über weitere Distanzen um als Deutsche und Menschen mit Migrationshintergrund. Darüber hinaus konnte festgestellt werden, dass Migranten mit gleichen Namen häufiger geografisch klumpen als es bei Deutschen und Menschen mit Migrationshintergrund der Fall ist.

Aufgrund der genannten Effekte muss erwartet werden, dass für Migranten eine schlechtere Linkage-Qualität bestehen wird als bei Deutschen oder Menschen mit Migrationshintergrund. Insbesondere Großgebäude und Großwohnsiedlungen können dabei Probleme bei der Verknüpfung verursachen. Aufgrund gleicher Adressen und den beobachteten Klumpen von Menschen mit gleichem Namen ist die Verwechslungswahrscheinlichkeit in diesen Gebieten wahrscheinlich besonders hoch. Diese Gebiete müssen im Zuge der Verknüpfung daher weiter untersucht werden.

Die Untersuchung der Fehler zeigt, dass die Dateneingaben der Schulen vielfach nicht den Angaben aus dem Melderegister entsprechen. Die Schulen erfassen häufiger nicht die zweite Staatsbürgerschaft und bei Vor- und Nachnamen häufig nur den ersten Token. Es muss beachtet werden, dass bei der Erstellung des ZSRs zunächst alle Personen von den Schulen gemeldet und dann mit dem Melderegister verknüpft werden. Dies bedeutet, dass im Folgenden keine Rohdaten verknüpft werden, sondern durch das Melderegister ergänzte Daten. Es ist erwartbar, dass die Linkage-Qualität mit den Rohdaten der Schulen deutlich schlechter ausfallen würde.

Im Vergleich zur Simulation des bundesweiten Bildungsverlaufsregisters (Schnell 2022) kann festgestellt werden, dass die simulierten Veränderungen nur zum Teil Änderungen im ZSR erklären. Die simulierten Fehlerquoten können nicht direkt mit den empirischen Ergebnissen verglichen werden, da sich empirisch keine konstante Fehlerquote für alle QIDs zeigt. So haben Namen eine deutlich höhere Fehlerquote als Geburtsdaten. Für eine Verknüpfung bedeutet dies, dass Blocking-Strategien, die auf Geburtsdaten aufbauen, zu empfehlen sind. Die Annahme, dass die Daten von Migranten eine höhere Fehlerquote als die Daten von Einheimischen haben, kann bestätigt werden. Jedoch ist der Faktor

meist größer als das Doppelte, so wie es in der Simulation des Bildungsverlaufsregisters angenommen wurde. Dies bedeutet, dass wahrscheinlich ein stärkerer Linkage-Bias nach Herkunft zu beobachten ist, als in der Simulation festgestellt wurde.

5. Datenverknüpfung

Aufbauend auf den bisherigen gesammelten Ergebnissen erfolgt im folgenden Kapitel die Verknüpfung der Schuljahre 2021/22 und 2023/24 des ZSR. Hierzu werden zunächst die verwendeten Record-Linkage-Methoden und getesteten Linkage-Szenarien beschrieben. Als Linkage-Szenario wird im Folgenden eine Kombination an QIDs bezeichnet, die für die Datenverknüpfung zur Verfügung stehen. Anhand der Linkage-Qualität einer Vielzahl von Szenarien kann die Relevanz (Datenqualitätsdimension) einzelner QIDs für ein Bildungsverlaufsregister abgeschätzt werden. Die getesteten Szenarien sollen darüber hinaus den Anspruch erheben, Referenzwerte für die Verknüpfung von amtlichen Datensätzen mit verschiedenen Record-Linkage-Verfahren und zur Verfügung stehenden QIDs zu geben.

Insgesamt wurden 40 Szenarien getestet und 390 Verknüpfungen durchgeführt. Eine detaillierte Analyse aller Szenarien kann im Rahmen dieser Arbeit nicht geleistet werden. Daher erfolgt im weiteren Verlauf eine Fokussierung auf drei ausgewählte Szenarien. Die Auswahl der Szenarien wird in Abschnitt 5.3 beschrieben.

5.1. Record-Linkage-Methoden

Im folgenden Abschnitt werden die verwendeten Record-Linkage-Methoden, die genaue Implementation und die Parameterwahl diskutiert. Das Vorgehen und die Methode ähnelt zur Vergleichbarkeit größtenteils der Simulation des Bildungsverlaufsregisters (Schnell 2022; Weiand 2022). Zusätzlich wurde als neue und bisher nicht getestete Methode das Verfahren von Datavant implementiert.

Die Verknüpfung erfolgte bei allen Verfahren zweistufig. Zunächst wurden alle Records als Match klassifiziert, die auf allen gegebenen QIDs exakt übereinstimmen (exakter Matchkey). Diese Records wurden gefiltert, sodass das zweite Verfahren nur die Restmenge verknüpft. Für alle Verfahren gilt, dass nur eindeutige Matches klassifiziert wurden. Hat ein Verfahren mit gleicher Sicherheit mehrere Paare zu einem Record als Match identifiziert, so wurde keins dieser Paare als Match klassifiziert. Ein Clerical-Review wurde nicht durchgeführt.

5.1.1. Probabilistisches Record-Linkage mit ECM

Probabilistisches Record-Linkage baut auf dem von Fellegi & Sunter (1969) vorgestellten Modell zur Berechnung einer Match-Wahrscheinlichkeit auf. In der großen Mehrheit aller Studien zum Vergleich von Record-Linkage-Methoden liefert probabilistisches Record-Linkage die beste Linkage-Qualität (Herzog et al. 2007; Christen 2012; Campbell 2009). Dies war auch bei der Simulation der Fall (Weiand 2022). Die Methode soll an dieser Stelle nur in den wesentlichen Grundzügen erläutert werden. Details finden sich unter anderem bei Herzog et al. (2007).

Zentral für das Modell sind die m- und u-Wahrscheinlichkeiten. Gegeben der zu verknüpfenden Datensätze A und B können zunächst die Mengen M und U definiert werden als:

$$M = \{(a, b); a \in A, b \in B, \text{(a,b) ist ein true Match}\} \tag{5.1}$$

und

$$U = \{(a, b); a \in A, b \in B, \text{(a,b) ist } \textbf{kein} \text{ true Match}\}. \tag{5.2}$$

Für jede QID i ist die m- und u-Wahrscheinlichkeit definiert als

$$m_i = \{a_i = b_i | (a, b) \in M; a \in A, b \in B\} \tag{5.3}$$

und

$$u_i = \{a_i = b_i | (a, b) \in U; a \in A, b \in B\}. \tag{5.4}$$

Die Wahrscheinlichkeit, dass zwei gleiche Ausprägungen einer QID zur gleichen Entität (Person) gehören, wird durch m_i ausgedrückt. Die Wahrscheinlichkeit, dass zwei gleiche Ausprägungen zu verschiedenen Entitäten gehören, wird durch u_i ausgedrückt. Die m- und u-Wahrscheinlichkeiten können über den EM-Algorithmus (Expectation Maximization) direkt aus den Daten bestimmt werden (**Winkler2000**).

Beim Vergleich zweier Records kann anhand der m- und u-Wahrscheinlichkeiten für jede QID ein Gewicht bestimmt werden. Abhängig davon, ob die betrachtete QID in beiden Records übereinstimmt oder nicht, ist das Gewicht w_i definiert als:

$$w_i = \begin{cases} \ln\left(\frac{m_i}{u_i}\right) & \text{falls die Merkmale übereinstimmen} \\ \ln\left(\frac{1-m_i}{1-u_i}\right) & \text{falls die Merkmale nicht übereinstimmen.} \end{cases} \tag{5.5}$$

Die Summe der Gewichte kann dann in eine Match-Wahrscheinlichkeit für ein Record-Paar überführt werden. Für alle Szenarien wurde ein Schwellenwert für die Match-Wahrscheinlichkeit von 0.8 und 0.9 getestet.

Ein großes Problem beim probabilistischen Record-Linkage besteht darin, dass zumeist sehr viele Record-Paare verglichen werden müssen. Dabei sind die Vergleiche rechenaufwendig. Eine übliche Methode zur Reduktion der Anzahl-Vergleiche sind *Blocking-Regeln* (Christen 2012: 69). Hierbei werden nur Record-Paare miteinander verglichen, die nach einer Regel übereinstimmen.

Zur Vergleichbarkeit der Ergebnisse des probabilistischen Record-Linkage wurde für alle Szenarien die gleiche zweistufige Blocking-Strategie implementiert. Die Strategie sieht vor, dass zunächst alle Records nach einer ersten Blocking-Regel verknüpft werden. Alle Records, die durch die erste Regel nicht als Match klassifiziert wurden, werden durch eine zweite Blocking-Regel erneut verknüpft. Für die Verknüpfung des ZSR wurden die Blocking-Regeln:

1. Exakte Übereinstimmung Geburtsmonat und Geburtsjahr und
2. exakte Übereinstimmung Soundex des Vornamens

verwendet. Die Parameter wurden auf Basis einer Zufallsstichprobe ($n = 50,000$) aus den vollständigen Datensätzen für beide Blocking-Verfahren einzeln geschätzt. Die Stichprobe erfolgte immer mit dem gleichen Startwert für den Pseudo-Zufallszahlengenerator, sodass die Parameter aller Szenarien auf Basis der gleichen Stichprobe berechnet wurden. Als Startwerte für den EM wurde das von Jaro (1989) vorgeschlagene Verfahren verwendet.

Für die Verknüpfung und Parameterberechnung wurde das Paket `recordlinkage` (de Bruin 2015; Version 0.16) verwendet. Das gleiche Paket wurde bei der Simulation verwendet. Die Ergebnisse sind somit vergleichbar (Schnell 2022). Das Paket implementiert anstatt des EM eine von Meng & Rubin (1993) präsentierte Optimierung des Algorithmus, den ECM (Expectation Conditional Maximization). Im Folgenden wird daher der Begriff ECM verwendet.

Die Vorteile des Pakets liegen in der leichten Implementierbarkeit von komplexen Vergleichsverfahren sowie der guten Skalierbarkeit. Der größte Nachteil des Pakets besteht darin, dass das Fellegi-Sunter-Modell nur in der Grundform mit einer binären Übereinstimmung von Merkmalen implementiert ist. So beschreiben Enamorado et al. (2019) eine Modifikation des Fellegi-Sunter-Modells, bei der durch die Einführung mehrerer Übereinstimmungsebenen ein partielles Übereinstimmen gewichtet werden kann. Dies führt häufig zu leicht besseren Ergebnissen (Enamorado et al. 2019).

Im Rahmen dieser Arbeit soll ein Fokus auf die Optimierung des probabilistischen Record-Linkage anhand der Ergebnisse aus Abschnitt 4.1 gelegt werden. Eine solche Optimierung war nur durch die Verwendung des Pakets `recordlinkage` möglich, da die Vergleichsverfahren leicht implementiert werden konnten und zudem bereits Erfahrung im Umgang bestand. Daher wurde auf die Verwendung von Paketen der von Enamorado et al. (2019) vorgeschlagenen Optimierungen (`splink` und `fastLink`) verzichtet.

Sofern die Merkmale Vorname, Nachname oder Straße vorlagen, wurden diese anhand der Jaro-Winkler-Ähnlichkeit klassifiziert (Winkler 1990). Alle anderen Merkmale wurden als übereinstimmend klassifiziert, wenn sie sich exakt gleichen. Für die Jaro-Winkler-Ähnlichkeit wurde der Schwellenwert 0.9 verwendet.

Aus den Ergebnissen von Abschnitt 4.1 lassen sich zwei Probleme identifizieren:

1. Die Veränderung der Adresse durch Umzüge und
2. die Einbeziehung unterschiedlicher Mengen an Token in Vor- und Nachname.

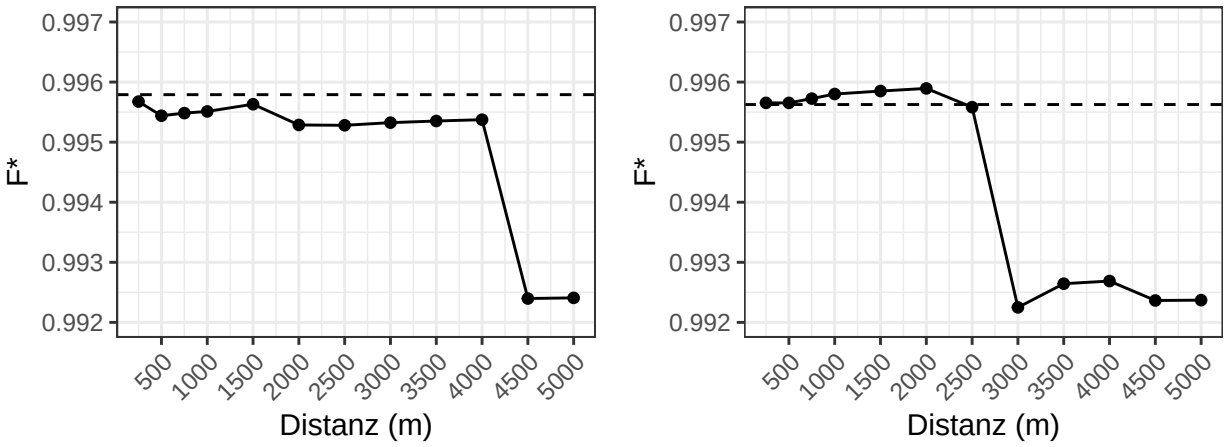
Für jedes Problem wurde eine Optimierung getestet.

Da die meisten betrachteten Personen auf kurzen Distanzen umziehen, kann die Veränderung der Adresse durch die Einbeziehung der Distanz zweier Adressen berücksichtigt werden. Das verwendete Paket erlaubt wie beschrieben nicht die Einbeziehung einer kontinuierlichen Variablen. Dementsprechend wurde nach einem Schwellenwert für eine maximale Distanz zweier Adressen gesucht. Die Optimierung kann zudem nur angewendet werden, falls eine geografische Koordinate vorliegt. Für die Suche nach einem geeigneten Schwellenwert wurde die 100m-Gitter-ID verwendet.

Abbildung 5.1 stellt die Linkage-Ergebnisse für zwei Szenarien mit unterschiedlichen Schwellenwerten dar. Es zeigt sich, dass die Berücksichtigung der Distanz nur in einem Szenario bessere Ergebnisse produziert als eine exakte Übereinstimmung mit der Gitter-ID. Der Kurvenverlauf deutet darauf hin, dass die geografische Distanz in einer weiteren Übereinstimmungsebene berücksichtigt werden sollte. Geeignete Schwellenwerte für den Radius wären 2 und 4 km. Da sich die Verbesserung nicht konsistent zeigte und eine andere, im Rahmen dieser Arbeit nicht umsetzbare Implementierung geeigneter für den Umgang mit Distanzen wäre, wurde auf die weitere Verwendung dieser Optimierung verzichtet. Die Ergebnisse deuteten jedoch eindeutig darauf hin, dass eine Berechnung der geografischen Distanz ein Verbesserungspotenzial für das Record-Linkage darstellt.

Die Analyse der Token (Abschnitt 4.5.2) hat gezeigt, dass die meisten Unterschiede zwischen Namen nicht auf Tippfehler, sondern auf eine unterschiedliche Menge verwendeter Token zurückzuführen sind. Bei einem direkten Vergleich der Namen über die Jaro-Winkler-Ähnlichkeit würde eine sehr geringe Ähnlichkeit zwischen den Namen festgestellt werden, falls einige Token fehlen. Obwohl die Jaro-Winkler-Ähnlichkeit in den meisten Fällen die besten Linkage-Ergebnisse bei der Verknüpfung von Namen erzielt, so gilt es anzumerken, dass diese für den Vergleich von einzelnen Namen entwickelt wurde und nicht für Mehrfachnamen (Christen 2012: 109; Yancey 2005).

Aus dem beschriebenen Problem ergibt sich, dass die Ähnlichkeit der Menge an Namen bestimmt werden muss. Hier eignen sich Ähnlichkeitsmaße für Mengen, wie z. B. der Dice-Koeffizient (Christen



(a) Vorname, Nachname, Geburtstag, -monat, -jahr, Geschlecht, 100m-Gitter-ID, Staatsangehörigkeit (Szenario 3)

(b) Vorname, Nachname, Geburtstag, -monat, -jahr, Geschlecht, 100m-Gitter-ID (Szenario 8)

Abbildung 5.1.: Linkage-Ergebnisse (F^*) bei Verwendung einer Distanz zur Klassifizierung der Übereinstimmung von Adressen für zwei Szenarien. Die gestrichelte Linie stellt das Ergebnis dar, wenn die 100m-Gitter-ID exakt übereinstimmen muss. In Klammern wird für eine bessere Vergleichbarkeit die Nummer später verwendeten Linkage-Szenarios angegeben (Abschnitt 5.2).

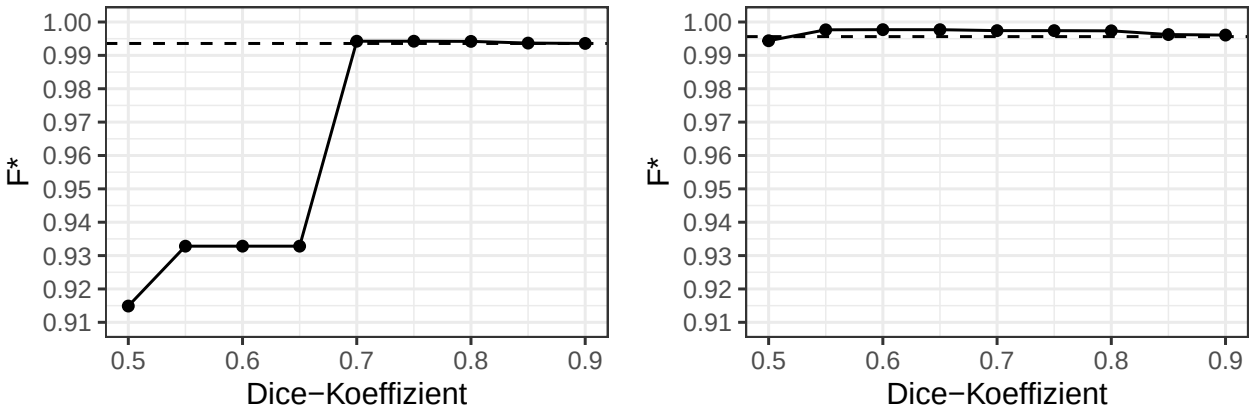
2012: 107). Dieser gibt die Ähnlichkeit zwischen zwei Mengen A und B als

$$sim_{Dice}(A, B) = \frac{2 \cdot |A \cap B|}{|A| + |B|} \tag{5.6}$$

an.

Zwei Namen stimmen nach dieser Methode überein, falls der Dice-Koeffizient der Token größer ist als ein vorher festgelegter Schwellenwert. Bei der Bestimmung der Schnittmenge kann die Jaro-Winkler-Ähnlichkeit verwendet werden. Zwei Token stimmen überein, falls die Jaro-Winkler-Ähnlichkeit größer ist als ein weiterer vorher festgelegter Schwellenwert. Das Vorgehen führt dazu, dass die Jaro-Winkler-Ähnlichkeit nur auf einzelne Namen angewendet wird. Die Methode ermöglicht darüber hinaus Vor- und Nachname nicht getrennt, sondern als Blockname auszuwerten. Hierdurch können auch Vertauschungen von Vor- und Nachname sowie unterschiedliche Platzierungen von Token im Vor- und Nachnamen berücksichtigt werden. Für die Berechnung werden somit zunächst die beiden Blocknamen in jeweils die Mengen einzigartiger Tokens aufgeteilt (A und B). Zur Bestimmung der Menge $A \cap B$ wird jedes Token in A wird mit B verglichen. Falls die Jaro-Winkler-Ähnlichkeit des Vergleichs größer als der festgelegte Schwellenwert ist, wird das Token in die Menge $A \cap B$ aufgenommen.

Abbildung 5.2 stellt die Ergebnisse für zwei Szenarien mit unterschiedlichen Schwellenwerten für den Dice-Koeffizienten dar. Getestet wurden jeweils Blocknamen und eine Jaro-Winkler-Ähnlichkeit der Token von 0.9. Die Ergebnisse zeigen, dass die Methode zum Teil leicht bessere Ergebnisse produziert. Diese Modifikation wurde daher – neben dem unmodifizierten Verfahren – als weiteres Record-Linkage-Verfahren bei allen Linkage-Szenarien angewendet, sofern Vor- und Nachname vorlagen. In diesen Fällen wurde anstelle der getrennten Namen der Blockname verwendet. Als Schwellenwert für den Dice-Koeffizienten wurde 0.75 und als Schwellenwert der Jaro-Winkler-Ähnlichkeit 0.9 verwendet. Das



(a) Blockname, Geburtstag, -monat, -jahr, Geschlecht, PLZ, Straße, Hausnummer, Staatsangehörigkeit (Szenario 1)

(b) Blockname, Geburtstag, -monat, -jahr, Geschlecht, 100m-Gitter-ID (Szenario 8)

Abbildung 5.2.: Linkage-Ergebnisse (F^*) bei Verwendung des Dice-Koeffizienten zur Klassifizierung der Übereinstimmung von Blocknamen für zwei Szenarien. Die gestrichelte Linie stellt das Ergebnis des unmodifizierten Verfahrens dar. In Klammern wird für eine bessere Vergleichbarkeit die Nummer später verwendeten Linkage-Szenarios angegeben (Abschnitt 5.2).

unmodifizierte Verfahren wird im Folgenden auch als ECM abgekürzt und das modifizierte Verfahren als ECM(D).

5.1.2. Cryptographic Longterm Keys mit Multibit-Trees

Cryptographic Longterm Keys (CLKs) sind eine PPRL-Methode, die auf Bloom-Filter basieren (Schnell et al. 2009; Schnell et al. 2011). Ein Bloom-Filter ist ein Bit-Vektor mit einer Länge l, bei dem die Elemente einer Menge (z. B. Bigramme eines Namens) durch k verschiedene Hash-Funktionen auf den Vektor abgebildet werden. Anhand der Ausprägungen der Bits kann eine Ähnlichkeit zwischen Bloom-Filtern bestimmt werden.

Ein CLK ist ein sehr langer Bloom-Filter, bei dem alle QIDs in einen Vektor eingetragen werden. Schnell et al. (2011) verwenden hierfür z. B. Bloom-Filter mit $l = 1000$ und $k = 10$. Die Speicherung einer Vielzahl von QIDs in einem einzigen Bit-Vektor ermöglicht die Verwendung von Multibit-Trees zur effizienten Suche von Bit-Vektoren (Kristensen et al. 2010; Brown et al. 2017; Borgs 2019). Diese wurden auch bei der Simulation verwendet (Schnell 2022).

Für die Verknüpfung des ZSRs konnte der Vergleich von CLKs durch Multibit-Trees ohne eine weitere Blocking-Regel durchgeführt werden.[1] Zur Bestimmung der Ähnlichkeit zwischen zwei CLKs wurde der Tanimoto-Koeffizient verwendet. Record-Paare mit einem Koeffizienten über einem Schwellenwert wurden als Match klassifiziert. Sollten mehrere Paare des gleichen Records über dem Schwellenwert liegen, so wurde das Paar mit dem höchsten Tanimoto-Koeffizienten gewählt. Für alle Szenarien wurden die Schwellenwerte 0.8 und 0.85 getestet. Vorname, Nachname und Straße wurden als Bigramm in den CLK eingetragen, alle anderen Merkmale als Unigramm. Es wurden CLKs mit $l = 1000$ und

[1] Folgendes Paket wurde verwendet: https://github.com/germanrecordlinkage/multibitTree (Version 1.7; abgerufen am 07.02.2024).

$k = 10$ verwendet. Die CLKs wurden mit der entsprechenden Methode aus dem R-Paket `PPRL` (Version 0.3.8) generiert.

Die Verwendung einer gleichen Anzahl Hash-Funktionen für alle Merkmale ist nicht optimal. Für ein Merkmal mit einer Menge an abzubildenden Elementen M, beträgt die maximale Anzahl Bits, die durch das Merkmal auf 1 gesetzt werden können $n_1 = |M| \cdot k$. Dies bedeutet, dass eine indirekte Gewichtung durch die Anzahl Elemente eines Merkmales erfolgt. Ein Ausgleich der Gewichtung ist über die Anpassung des Parameters k für das Merkmal möglich. Da es zurzeit kein Verfahren zur Gewichtung bei CLKs gibt und dieses im Rahmen dieser Arbeit nicht entwickelt werden kann, musste die gleiche Anzahl Hash-Funktionen für alle QIDs verwendet werden.

Weitere Probleme entstehen bei der Verwendung von Adressdaten. So können Adressdaten dazu führen, dass nicht umgezogene Personen besser verknüpfbar werden, während umgezogene Personen schlechter verknüpfbar werden. Die Verwendung von Adressdaten kann somit einen Linkage-Bias einführen. Insbesondere bei der Verwendung der vollen Anschrift ist ein Linkage-Bias wahrscheinlich, da sich bei einem Umzug viele Bits ändern. Ein entsprechender Umgang mit Umzügen wäre für die CLKs daher auch notwendig. Insbesondere die Einbeziehung der Entfernung zwischen zwei Koordinaten könnte die Linkage-Qualität verbessern.

Die Berechnung einer approximativen Distanz zwischen zwei Koordinaten ist auch verschlüsselt möglich. So stellen Schnell et al. (2021) ein Verfahren vor, bei der zunächst eine Menge Zufallszahlen in einem festen Schema einer Geo-Koordinate zugeordnet werden (z. B. als Gitter). Für jedes Record wird dann die Menge Zufallszahlen gespeichert, die in einem vorher festgelegten Radius liegen. Anhand der Menge übereinstimmender Zufallszahlen zwischen zwei Records kann schließlich eine approximative geografische Distanz berechnet werden.

Die Anwendung dieser Methoden ist im Rahmen dieser Arbeit nicht möglich. Ein Hauptproblem stellt dabei die Gewichtung der Distanz dar. Eine Annäherung kann jedoch durch die Verwendung einer Gitter-ID erreicht werden. Je näher zwei Records geografisch aneinander liegen, desto mehr Dezimalstellen der Gitter-IDs gleichen sich. Damit steigt die Ähnlichkeit der beiden Records. Die Leistungsfähigkeit der Gitter-ID kann daher weiteren Aufschluss über den Nutzen distanzbasierter Verfahren bringen.

5.1.3. Multiple Matchkeys

Randall et al. (2019) stellen eine Methode vor, bei der eine Vielzahl an deterministischen Matchkeys verwendet werden, um ein Match zu klassifizieren. Das Verfahren zeigte in der Simulation sehr gute Ergebnisse. Es wurde jedoch die Hypothese aufgestellt, dass sich diese guten Ergebnisse nur zeigten, weil die simulierten Datensätze immer eine Obermenge des vorherigen Jahrs waren (Weiand 2022). Diese Hypothese kann im Folgenden überprüft werden, da die Datensätze des ZSR nur in einer Schnittmenge miteinander übereinstimmt.

Die Matchkeys werden auf Basis der Gewichte aus dem Fellegi-Sunter-Modell gebildet. Die Gewichte können dabei über den EM-Algorithmus bestimmt werden. Hierzu werden zunächst alle Kombinationen von QIDs, deren summierte Gewichte über einem Schwellenwert liegen, als mögliche Matchkeys identifiziert. Die einzelnen Gewichte stellen jeweils Logits dar. Durch das Einsetzen in eine logistische Funktion lässt sich daher die Summe der Gewichte R in eine besser interpretierbare Wahrscheinlichkeit umrechnen (Linacre 2023):

$$p = \frac{e^R}{1 + e^R}. \tag{5.7}$$

In der Simulation zeigte sich, dass niedrige Schwellenwerte die besten Ergebnisse produzieren (Weiand 2022). Da es sich um eine andere Datengrundlage handelt, wurden für alle Szenarien die Schwellenwerte 0.7, 0.9 und 0.98 getestet.[1] Die Parameter wurden für alle Schwellenwerte auf Basis einer Zufallsstichprobe ($n = 7,000$) aus den vollständigen Datensätzen berechnet, ohne Verwendung einer Blocking-Regel. Die Berechnung erfolgte mit dem Paket `recordlinkage` (de Bruin 2015). Die Stichproben wurden immer mit den gleichen Startwerten für den Pseudo-Zufallszahlengenerator gezogen.

Nachdem alle Matchkeys über dem Schwellenwert identifiziert wurden, sieht die Methode einen Ausschluss von redundanten Matchkeys vor. Ein Matchkey ist redundant, wenn ein weiterer Matchkey vorliegt, der aus einer exakten Teilmenge der Merkmale des ersten Matchkeys besteht. Ein Match entsteht demnach unabhängig, ob das Merkmal übereinstimmt oder nicht. Die so reduzierte Anzahl Matchkeys wird verwendet, um die Datensätze zu verknüpfen.[2]

Da Randall et al. (2019) keine Angaben über die Klassifikation eines Matches geben, wurde die bereits in der Simulation verwendete Implementation übernommen (Weiand 2022). Hierbei wird die Anzahl übereinstimmender Matchkeys für ein Record-Paar bestimmt. Das Paar mit der höchsten Anzahl übereinstimmender Matchkeys für beide Records wird als Match klassifiziert. Liegen gleiche Häufigkeiten vor, so wird gemäß der zu Beginn aufgestellten Regeln kein Match klassifiziert.[3]

5.1.4. Datavant

Datavant ist ein Unternehmen, welches im Bereich der Verknüpfung von Gesundheitsdaten tätig ist.[4] Eine genaue Beschreibung der verwendeten Record-Linkage-Methode ist nicht öffentlich zugänglich. Aus einigen bekannten Details konnte jedoch ein Verfahren abgeleitet werden, welches den Ergebnissen des Datavant-Verfahrens ähneln sollte. Das hier präsentierte Verfahren ähnelt zudem auch einem vom Office for National Statistics (ONS) vorgestellten Verfahren, welches auch auf die Verwendung einer Vielzahl vordefinierter Matchkeys basiert (Shipsey & Plachta 2021).

Für die Verknüpfung nutzt Datavant eine Menge an Matchkeys, die aus den zur Verfügung stehenden Merkmalen gebildet werden können (Datavant 2020). Diese Matchkeys werden anschließend über eine Hash-Funktion und einem Salt verschlüsselt, sodass es sich um ein PPRL-Verfahren handelt (Gupta et al. 2023; Mirel et al. 2022). Da Kollisionen durch gleiche Hash-Werte höchst unwahrscheinlich sind, wurde auf eine Verschlüsselung der Ergebnisse verzichtet. Die Verknüpfung erfolgt durch exakten Abgleich der unverschlüsselten Matchkeys. Die Matchkeys werden zudem unterschiedlich gewichtet:

"The weighting algorithm is executed to assign weights to hashes included in a given set of hashes based on the characteristics of the population attributes on which the weights are being applied. [...] In some embodiments, hash's weights remains the same within the project and don't change to allow consistent matching to occur" (Kho & Goel 2019: 6f.).

[1] Ursprünglich wurden auch die Schwellenwerte 0.8 und 0.95 getestet. Die Darstellung dieser Ergebnisse ergaben jedoch keinen Mehrwert, sodass diese Werte enfallen sind.

[2] Die Implementation beruht auf dem Originalcode von Randall et al. (2019). Aus Gründen der Performance wurde der Algorithmus zur Bestimmung von redundanten Matchkeys vollständig überarbeitet. Beide Algorithmen produzieren die gleichen Ergebnisse.

[3] Technisch wurde die Zählung der übereinstimmenden Matchkeys über einen join-Befehl gelöst. In einem Fall kam es dazu, dass ein Matchkey zu wenig mögliche Kombinationen besaß, sodass die Datensätze auf dem verwendeten Rechnersystem nicht mehr zusammengeführt werden konnten (Matchkey: Geschlecht, Staatsangehörigkeit). In diesem Fall wurden aus den bestehenden berechenbaren Matchkeys und dem nicht berechenbaren Matchkey neue Matchkeys erzeugt und für das Record-Linkage verwendet. Hierzu wurde jedem berechenbaren Matchkey der nicht berechenbare Matchkey angefügt. Der nicht berechenbare Matchkey wurde dann nicht weiter für die Verknüpfung verwendet.

[4] https://datavant.com/ (abgerufen am 09.03.2023).

Nr.	Matchkey
1	1. Buchstabe Vorname, Nachname, Tag, Monat, Jahr, Geschlecht
2	Soundex Vorname, Soundex Nachname, Tag, Monat, Jahr, Geschlecht
3	Blockname, Tag, Monat, Jahr, 1.-2. Stelle PLZ
4	Blockname, Tag, Monat, Jahr, Geschlecht
5	1.-3. Buchstabe Vorname, Nachname, Tag, Monat, Jahr, Geschlecht
6	Vorname, Hausnummer, Straße, PLZ
7	Vor- und Nachname vertauscht, Tag, Monat, Jahr, Geschlecht
8	Soundex Vorname, Nachname, Tag, Monat, Jahr
9	Soundex Vorname, Soundex Nachname, Jahr, Hausnummer, Straße, PLZ
10	Blockname, Monat, Staatsbürgerschaft, Hausnummer, Straße, PLZ
11	Nachname, Tag, Monat, Jahr, Geschlecht, Staatsbürgerschaft, Straße, PLZ
12	Blockname, Tag, Jahr, Hausnummer, Straße
13	1.-3. Buchstabe Vorname, Nachname, Monat, Jahr, 100m-Gitter-ID
14	1.-3. Buchstabe Vorname, Nachname, Tag, Monat, 100m-Gitter-ID
15	1.-3. Buchstabe Vorname, Tag, Monat, Jahr, 100m-Gitter-ID
16	Nachname, Tag, Monat, Jahr, Geschlecht, 100m-Gitter-ID
17	Blockname, Monat, Jahr, 1km-Gitter-ID
18	Blockname, Tag, Monat, 1km-Gitter-ID
19	Vorname, Soundex Nachname, Tag, Monat, Jahr, 1km-Gitter-ID
20	Nachname, Tag, Monat, Jahr, Geschlecht, 1km-Gitter-ID
21	Vorname, Soundex Nachname, Monat, Jahr, Staatsbürgerschaft

Tabelle 5.1.: Verwendete Matchkeys für das Datavant-Verfahren.

Da keine weiteren Details zum Gewichtungs-Algorithmus veröffentlicht wurden, erhielten alle Matchkeys das gleiche Gewicht. Auch zur Klassifikation der Matches wurden keine Details publiziert. Daher wurde der Klassifikator aus dem Multiple-Matchkey-Verfahren übernommen. Die Klassifikation erfolgte somit anhand der Zählung übereinstimmender Matchkeys. Liegen mehrere Record-Paare mit einer gleichen Anzahl übereinstimmender Matchkeys vor, erfolgte keine Verknüpfung. Zur Klassifikation eines Record-Paars genügt ein übereinstimmender Matchkey.

Eine vollständige Liste der von Datavant verwendeten Matchkeys ist nicht öffentlich. Bernstam et al. (2022) zeigen jedoch eine Reihe an Matchkeys, die das Unternehmen unter anderem verwendet. Auf Grundlage dieser Publikation wurden 21 verschiedene Matchkeys manuell zusammengestellt (siehe Tabelle 5.1). Matchkey 1-6 konnten aus Bernstam et al. (2022) übernommen werden. Zur Zusammenstellung der anderen Matchkeys wurden drei Prämissen aufgestellt:

1. Es darf keine indirekte Gewichtung erfolgen,
2. eine Klassifikation kann durch einen einzigen Matchkey erfolgen und
3. für jedes Szenario soll eine ausreichende Anzahl Matchkeys zur Verfügung stehen.

Das größte Problem bei der Zusammenstellung der Matchkeys stellt eine indirekte Gewichtung der Merkmale dar. Eine indirekte Gewichtung entsteht über die Häufigkeit eines Merkmals oder einer Merkmalskombination in allen verwendeten Matchkeys. Falls das wahre Record-Paar nicht in diesen Merkmalen übereinstimmt, entsteht eine niedrigere Summe an übereinstimmenden Matchkeys. Falls zu einem dieser Records ein anderer Record vorliegt, bei dem diese Merkmale übereinstimmen, wird dieses Paar höher gewichtet. Eine falsche indirekte Gewichtung führt somit zu mehr falsch Positiven.

Zur Klassifikation eines Record-Paars ist es möglich, einen Schwellenwert für eine Mindestanzahl übereinstimmender Matchkeys festzulegen. In diesem Fall können Matchkeys definiert werden, die Personen nicht eindeutig identifizieren müssen. Das Zusammenspiel solcher Matchkeys, insbesondere durch die entstehende indirekte Gewichtung, führt jedoch zu großen Problemen bei der Auswahl der Matchkeys. Entsprechend wurde kein Schwellenwert eingeführt. Dies erleichterte die Auswahl der Matchkeys. Jeder Matchkey kann potenziell ein Record-Paar verknüpfen und darf mit großer Sicherheit keine falsch positiven Verknüpfungen erzeugen.

Auch aufgrund der ersten beiden Prämissen konnten nicht alle Szenarien berücksichtigt werden. Der Fokus wurde daher auf Szenarien gelegt, in denen Vor- und Nachname vollständig verfügbar sind. Für die dritte Prämisse wurde insbesondere berücksichtigt, dass die getesteten Szenarien entweder eine Gitter-ID oder Adressangaben enthalten. Für keine Verknüpfung wurden daher alle 21 Matchkeys verwendet. Für die getesteten Szenarien sind maximal 13 Matchkeys simultan möglich. Die Anzahl verwendeter Matchkeys werden in den Linkage-Ergebnissen angegeben.

Bei der Auswahl der Matchkeys wurde zudem der Vorteil genutzt, dass auch Vertauschungen leicht berücksichtigt werden können (Matchkey 7). Des Weiteren konnte anstatt des Abgleichs von Vor- und Nachname der Blockname verwendet werden, sodass auch Personen mit einem Blocknamen verknüpft werden können.

5.2. Linkage-Szenarien

Bei der Zusammenstellung der Linkage-Szenarien wurde darauf verzichtet, alle möglichen Kombinationen an QIDs zu testen. Die vollständige Evaluation aller Szenarien ist im Rahmen dieser Arbeit weder möglich noch zielführend. Ferner würde eine Evaluation aller Szenarien unterschiedliche Record-Linkage-Strategien benötigen. Diese würde die Ergebnisse schlechter vergleichbar machen. Die betrachteten Szenarien wurden daher so gewählt, dass alle Szenarien mit der gleichen Strategie verknüpft werden können und eine Vergleichbarkeit weitestgehend erhalten bleibt.

Die getesteten Szenarien setzen sich aus einer Kombination an möglichen Variationen zusammen. Verwendet werden nur QIDs aus dem ZSR sowie die 100m- und 1km-Gitter-IDs. Eine wesentliche Variation ist das verfügbare Adressmerkmal. Für die einzelnen Szenarien wurde jeweils getestet, dass die

1. vollständige Anschrift (Straße, Hausnummer, Postleitzahl),
2. Postleitzahl,
3. 100m-Gitter-ID,
4. 1km-Gitter-ID oder
5. kein Adressmerkmal

vorliegt.

Als Zweites wurde die Verfügbarkeit der Staatsangehörigkeit variiert. Es wurde immer die zusammengefasste Staatsangehörigkeit verwendet (erste nicht deutsche Staatsangehörigkeit; Abschnitt 4.4.1). In der dritten Variation wurde angenommen, dass das Geburtsdatum ein datenschutzrechtlich bedenkliches Merkmal ist. Hierzu wurde das Vorhandensein des Geburtstags variiert. Das Vorhandensein des Geburtsmonats und des Geburtsjahres konnte nicht getestet werden, da diese Merkmale Teil der Blocking-Regeln des probabilistischen Record-Linkage sind. Der vollständige Name stellt ein weiteres datenschutzrechtlich bedenkliches Merkmal dar. In der letzten Variation wurde daher der anstatt des vollständigen Namens der Soundex des Vor- und Nachnamens verwendet.

Aus den möglichen Variationen ergaben sich insgesamt 40 Szenarien, die getestet wurden. Eine Auflistung aller getesteten Szenarien und zugehörigen QIDs findet sich in Tabelle 5.2.

5.3. Linkage-Qualität und -Bias aller Szenarien

Im Folgenden werden die Record-Linkage-Ergebnisse aller getesteter Szenarien zunächst beschrieben. Eine vollständige Auflistung aller Ergebnisse befindet sich in Tabelle A.1 im Anhang. Da die Parameter aller Verfahren auf die Maximierung der Precision ausgerichtet wurden, bestehen die größten Unterschiede zwischen den Verfahren im erreichten Recall. Um dennoch den Einfluss der false positives mitzuberücksichtigen, wird sich in der folgenden Darstellung vor allem auf das Gütemaß F^* bezogen.

Die Ergebnisse zeigen, dass in allen Szenarien eine gute bis sehr gute Linkage-Qualität erreicht wird. Wie zu erwarten, liegt die höchste Precision immer beim exakten Abgleich. Der Recall des exakten Abgleichs ist jedoch meistens am niedrigsten. In 36 der 40 Szenarien erzielt das probabilistische Record-Linkage den besten F^*-Wert. Davon erreicht die modifizierte Version in 16 Fällen den besten F^*-Wert. In 34 der 36 Fälle liegt der Schwellenwert für das beste Ergebnis des probabilistische Record-Linkage bei 0.8. Für beide Varianten des probabilistische Record-Linkage zeigt sich, dass eine Erhöhung des Schwellenwertes meist mit einer leichten Verbesserung der Precision einhergeht, jedoch mit einer deutlichen Verschlechterung des Recalls. Das höchste F^* produziert die modifizierte Variante des probabilistische Record-Linkage in Szenario 8 mit einem Schwellenwert von 0.8. In diesem Szenario werden 92 false positives und 380 false negatives klassifiziert.

Bei den restlichen vier Szenarien wird der höchste F^*-Wert in zwei Fällen durch das Datavant-Verfahren und in einem Fall durch die CLKs erreich. Die Linkage-Qualität des letzten Szenarios kann durch kein Verfahren verbessert werden, sodass der höchste F^*-Wert in diesem Fall beim exakten Verfahren liegt. Für die meisten Szenarien kann folgende absteigende Reihenfolge in den F^*-Werten beobachtet werden:

1. probabilistisches Record-Linkage,
2. Datavant,
3. CLK,
4. Multiple Matchkeys und
5. exakter Matchkey.

Insbesondere das Datavant-Verfahren stellt sich als sehr effektiv heraus. Das Verfahren produziert häufig eine sehr gute Linkage-Qualität, wobei die Verknüpfung der beiden Schuljahre in zumeist weniger als einer Sekunde erfolgt. Der höchste F^*-Wert wird im Szenario 3 erreicht (68 false positives, 605 false negatives), gefolgt von Szenario 8 (57 false positives, 626 false negatives). Bereits wenige Matchkeys reichen zudem aus, um eine sehr gute Linkage-Qualität zu erzielen. Ein Beispiel hierfür ist das Szenario 11. Im Vergleich zum exakten Verfahren klassifiziert das Datavant-Verfahren mit nur einem Matchkey $20,278$ weitere true positives und nur 51 false positives. Damit erreicht das Verfahren die beste Linkage-Qualität (F^* und Recall) des Szenarios.

Das Beispiel zeigt deutlich, dass gut gewählte Matchkeys eine wirksame unterstützende Methode für das Record-Linkage darstellen. Neben einer eigenständigen Verwendung von Matchkeys in einer Linkage-Strategie (z. B. als Filter), können Matchkeys durch eine parallele Anwendung mit einem anderen Record-Linkage-Verfahren als ein Auswahlkriterium für ein Clerical-Review verwendet werden. Letzteres soll anhand der Ergebnisse des unmodifizierten probabilistischen Record-Linkage und des Datavant-Verfahrens am Szenario 6 verdeutlicht werden:

Szenario	Vorname	Soundex(Vorname)	Nachname	Soundex(Nachname)	Geschlecht	Geburtstag	Geburtsmonat	Geburtsjahr	Postleitzahl	Straße	Hausnummer+Zusatz	Staatsangehörigkeit	100m-Gitter-ID	1km-Gitter-ID
1	X		X		X	X	X	X	X	X	X	X		
2	X		X		X	X	X	X	X			X		
3	X		X		X	X	X	X				X	X	
4	X		X		X	X	X	X				X		X
5	X		X		X	X	X	X				X		
6	X		X		X	X	X	X	X	X	X			
7	X		X		X	X	X	X	X					
8	X		X		X	X	X	X					X	
9	X		X		X	X	X	X						X
10	X		X		X	X	X	X						
11	X		X		X		X	X	X	X	X	X		
12	X		X		X		X	X	X			X		
13	X		X		X		X	X				X	X	
14	X		X		X		X	X				X		X
15	X		X		X		X	X				X		
16	X		X		X		X	X	X	X	X			
17	X		X		X		X	X	X					
18	X		X		X		X	X					X	
19	X		X		X		X	X						X
20	X		X		X		X	X						
21		X		X	X	X	X	X	X	X	X	X		
22		X		X	X	X	X	X	X			X		
23		X		X	X	X	X	X				X	X	
24		X		X	X	X	X	X				X		X
25		X		X	X	X	X	X				X		
26		X		X	X	X	X	X	X	X	X			
27		X		X	X	X	X	X	X					
28		X		X	X	X	X	X					X	
29		X		X	X	X	X	X						X
30		X		X	X	X	X	X						
31		X		X	X		X	X	X	X	X	X		
32		X		X	X		X	X	X			X		
33		X		X	X		X	X				X	X	
34		X		X	X		X	X				X		X
35		X		X	X		X	X				X		
36		X		X	X		X	X	X	X	X			
37		X		X	X		X	X	X					
38		X		X	X		X	X					X	
39		X		X	X		X	X						X
40		X		X	X		X	X						

Tabelle 5.2.: Zusammensetzung der getesteten Linkage-Szenarien.

Das Datavant-Verfahren klassifiziert in diesem Szenario 60 false positives und 702 false negatives. Das probabilistische Record-Linkage klassifiziert 85 false positives und 451 false negatives. Werden die Ergebnisse zusammengefasst, so werden 18 false positives von beiden Verfahren klassifiziert. Von 836 Record-Paaren, die von nur einem der beiden Verfahren als Match klassifiziert wurden, sind 124 false positives (14.8 %). Das Zusammenfassen der Verfahren führt somit zu einer effizienten Auswahl von problematischen Fällen, die im Rahmen eines Clerical-Reviews betrachtet werden sollten. In den zusammengefassten Ergebnissen liegen zudem nur noch 214 false negatives vor. Würde das Clerical-Review alle ausgewählten Paare richtig klassifizieren, würde dieses zusammengefasste Verfahren den besten F^*-Wert insgesamt produzieren. Ähnliche Ergebnisse zeigen sich bei den anderen Szenarien.

Im Gegensatz zu den Ergebnissen der Simulation sind die Ergebnisse der Multiple-Matchkey-Methode in allen Szenarien schlecht. Dies bestätigt die Hypothese, dass die Methode nur bei Datensätzen eine gute Linkage-Qualität liefert, bei der ein Datensatz eine Teilmenge des anderen Datensatzes darstellt (Weiand 2022). So klassifiziert die Methode häufig viele false positives. Dies kann darauf zurückgeführt werden, dass die Bereinigung von Redundanten Matchkeys sehr schlechte Matchkeys produziert. Das Szenario 10 liefert z. B. für alle Schwellenwerte die gleichen zwei reduzierten Matchkeys:

- Nachname, Geschlecht und
- Vorname.

Extrem schlechte Linkage-Ergebnisse werden zumeist nur dadurch verhindert, dass die Matchkeys nicht eindeutig eine Person identifizieren können. Es zeigt sich zudem, dass tendenziell eine höhere Precision bei einem hohen Schwellenwert erreicht wird und ein höherer Recall bei einem niedrigen Schwellenwert. Die Ergebnisse sind jedoch nicht konsistent. Die von Randall et al. (2019) vorgeschlagene Methode zur Reduktion der Matchkeys erzeugt daher nicht nur schlechte, sondern auch unzuverlässige Ergebnisse. Von einer Verwendung der Methode bei Datensätzen, die keine Teilmengen voneinander darstellen, wird abgeraten. Es muss jedoch hervorgehoben werden, dass die Reduktionsfunktion das Problem darstellt und nicht die Generierung von Matchkeys anhand der m- und u-Wahrscheinlichkeiten.

Die CLKs weisen sehr konsistente Linkage-Ergebnisse auf. Der höchste Recall liegt immer beim niedrigeren Schwellenwert und die höchste Precision beim höheren. In 24 der 40 Szenarien liegt der höhere F^*-Wert bei einem Schwellenwert von 0.8. Besonders gute Ergebnisse liefern die CLKs bei Szenarien, welche eine Gitter-Koordinate enthalten. So wird der höchste F^*-Wert bei Szenario 8 erreicht (165 false positves, $1,652$ false negatives). Dies deutet darauf hin, dass die Einbeziehung der geografischen Distanz die Linkage-Qualität verbessert. Die Verwendung der vollständigen Adresse führte hingegen zu einer erwarteten schlechten Linkage-Qualität, da Personen, die umgezogen sind, zumeist nicht verknüpft werden können.

Grundsätzlich führt die Verwendung der 100m-Gitter-ID zu besseren Linkage-Ergebnissen als die 1km-Gitter-ID. Für die meisten Verfahren bestehen zudem nur geringe Ergebnisunterschiede zwischen der Verwendung der Postleitzahl und der 1km-Gitter-ID. Die einzige Ausnahme stellen die CLKs dar, welche deutlich bessere Ergebnisse bei der Verwendung der 1km-Gitter-ID erzielen. Dies deutet erneut auf die möglichen Vorteile durch die Einbeziehung der geografischen Distanz hin.

Die Einbeziehung der Staatsbürgerschaft führt zu keiner eindeutigen Verbesserung, zum Teil sogar zu einer Verschlechterung der Ergebnisse. Daher ist die Staatsbürgerschaft kein relevantes Merkmal zur Verknüpfung eines bundesweiten Bildungsverlaufsregisters.

Im Vorfeld nicht erwartet wurde die hohe Linkage-Qualität bei Szenarien mit einer hohen Datensparsamkeit. So liegt der F^*-Wert für das Szenario 40 (Soundex Vorname, Soundex Nachname, Geburtsmonat, -jahr, Geschlecht) bei 0.9827. Unter Einbezug der 100m-Gitter-ID (Szenario 38) wird sogar ein F^*-Wert von 0.9944 erreicht. Zum Vergleich: Der insgesamt höchste F^*-Wert liegt bei 0.9974

(Szenario 8). Für kleinere Datensätze reichen diese Merkmale zur guten Klassifikation bereits aus. Wie oben beschrieben, ist dies für große Datensätze jedoch nicht zu erwarten. Insbesondere die exakte Verknüpfung des Szenarios 40 zeigt, dass 98.3 % der Records leicht zu verknüpfen sind.[1] Die Verknüpfung der übrigen 1.7 % ist hingegen mit einem hohen Aufwand verbunden. Kein getestetes Verfahren kann alle Records richtig verknüpfen. Die Modifikation des probabilistischen Record-Linkage, die guten Ergebnisse bei Einbeziehung der geografischen Distanz bei CLKs sowie die Kombination von Datavant und probabilistischen Record-Linkage zeigen, dass durchaus Verbesserungspotenzial in der Linkage-Qualität besteht. Es ist jedoch nicht zu erwarten, dass eine vollständige Verknüpfung des Datensatzes ohne ein Clerical-Review ausgewählter Fälle erreicht werden kann.

Für alle Szenarien wurde des Weiteren der Linkage-Bias in Form eines Effektstärkemaßes berechnet. Für metrische Variablen wurde Cohen's d verwendet, für nominale Variablen Cramer's V und für Anteilswerte Cohen's h (Cohen 1988). Es wurden jeweils die Werte zwischen dem wahren Datensatz und dem verknüpften Datensatz verglichen. Datenbasis war immer das Schuljahr 2021/22. Nach Cohen (1988) liegt ein schwacher Effekt für Cohen's d und Cohen's h bei einem Wert von 0.2 vor. Ein mittlerer Effekt bei einem Wert von 0.5 und ein starker Effekt bei einem Wert von 0.8. Ein schwacher Effekt für Cramer's V liegt bei 0.1 vor, ein mittlerer Effekt bei 0.3 und ein starker Effekt bei 0.5 (Cohen 1988). Effektstärken wurden für das Alter, die Schulformen, die Familiensprache, die Datenquellen, den Anteil der Migranten, Menschen mit Migrationshintergrund und Deutsche sowie den Geschlechteranteil und eine Kombination aus Herkunft und Geschlecht berechnet. Für kein Merkmal kann ein schwacher Effekt festgestellt werden (siehe Tabelle A.1 im Anhang). Dies bedeutet, dass kein Record-Linkage-Verfahren für die durchgeführte Verknüpfung einen univariaten Linkage-Bias erzeugt.[2]

Die Ergebnisse zeigen weiter, dass Adresswechsel bei einigen Verfahren zu Problemen führen. Hierzu wurde der Anteil umgezogener Personen berechnet, die durch das Verfahren verknüpft werden konnten.[3] Kein Verfahren kann alle umgezogenen Personen verknüpfen, wobei Verfahren mit besonders hohen Anteilen richtig verknüpfter umgezogener Personen meist eine niedrige Precision besitzen. Bei Verfahren mit einer insgesamt hohen Linkage-Qualität ($F^* > 0.995$) liegt der Anteil richtig verknüpfter umgezogener Personen meist zwischen 98 % und 99 %. Bei insgesamt $16,584$ umgezogenen Personen sind somit zwischen 166 und 332 false negatives auf nicht verknüpfte umgezogene Personen zurückzuführen. Dies bedeutet, dass bei den besten Record-Linkage-Verfahren im Mittel die Hälfte aller false negatives auf nicht verknüpfte umgezogene Personen zurückzuführen sind. Zwar können zumeist mehr umgezogene Personen bei Szenarien ohne Verwendung von Adressdaten richtig verknüpft werden, diese besitzen jedoch allgemein eine schlechtere Linkage-Qualität. Insbesondere das Datavant-Verfahren und das probabilistische Record-Linkage können auch bei vorliegenden Adressdaten umgezogene Personen zu einem hohen Anteil verknüpfen. CLKs erreichen dies nur bei Verwendung einer Gitter-Koordinate.

Die Ergebnisse zeigen, dass Maßnahmen zur Rückverfolgung von umgezogenen Personen in einem Bildungsverlaufsregister getroffen werden müssen. Eine Möglichkeit hierfür wäre ein historisiertes Anschriftenregister (Schnell 2019a: 39).

---

[1] Das exakte Verfahren verknüpft in Szenario 40 $179,265$ true positives. Bei $182,385$ wahren Matches ergeben sich $\frac{179,265}{182,385} = 0.983$.

[2] Aus Platzgründen wird auf die Darstellung der Effekte für die Familiensprache und das kombinierte Merkmal aus Herkunft und Geschlecht in der Ergebnistabelle verzichtet. Die Effekte für diese Merkmale sind besonders gering.

[3] Bei einer falsch positiven Verknüpfung mit einem Record, welches nur in einem Schuljahr des ZSRs enthalten ist, ist nicht bekannt, ob die Person umgezogen ist. Dieser Wert kann auch nicht imputiert werden, da der Wert MNAR ist (Missing Not At Random). Aus diesem Grund kann der Linkage-Bias für den Anteil umgezogener Personen nicht in Form einer Effektstärke berechnet werden.

5.4. Detailanalyse ausgewählter Linkage-Szenarien

Im folgenden Abschnitt werden die Linkage-Ergebnisse detailliert analysiert. Dabei werden insbesondere die Datenqualitätsdimensionen Genauigkeit und Zuverlässigkeit sowie Kohärenz und Vergleichbarkeit untersucht (Abschnitt 3.3). Eine detaillierte Analyse aller Linkage-Ergebnisse der vorgestellten 40 Szenarien ist im Rahmen dieser Arbeit weder möglich noch zielführend. Aus diesem Grund erfolgt zunächst eine Auswahl von drei Linkage-Szenarien, die mögliche und plausible Optionen für ein bundesweites Bildungsverlaufsregister darstellen. Darüber hinaus bedarf es einer Auswahl weniger Record-Linkage-Verfahren, die miteinander verglichen werden sollen. Die ausgewählten Szenarien und Verfahren werden auf Einflussfaktoren für eine Fehlverknüpfung sowie auf regionale Unterschiede geprüft.

5.4.1. Auswahl der analysierten Linkage-Szenarien und -Verfahren

Die analysierten Linkage-Szenarien wurden anhand der Ergebnisse aller Record-Linkage-Verfahren (Tabelle A.1) sowie der bisherigen Literatur zu einem bundesweiten Bildungsverlaufsregister ausgewählt. Die erste Grundlage war die Simulation des Bildungsverlaufsregisters. Hierbei wurden zwei Szenarien getestet (Schnell 2022):

1. Vorname, Nachname, Geburtstag, Geburtsmonat, Geburtsjahr, Geschlecht und
2. Vorname, Nachname, Geburtstag, Geburtsmonat, Geburtsjahr, Geschlecht, Geburtsort.

Die Ergebnisse der Simulation zeigten, dass das erste Szenario für bundesweites Bildungsverlaufsregister ungeeignet und der Geburtsort ein notwendiger Quasi-Identifikator ist. Da im ZSR keine Angaben über den Geburtsort enthalten sind, wird das erste Szenario als Szenario mit der höchsten Datensparsamkeit betrachtet (Minimalszenario). Das Minimalszenario entspricht dem Szenario 10. Die Wahl dieses Szenarios erlaubt zudem einen Vergleich mit den Ergebnissen der Simulation. Ferner ist zu erwarten, dass das Merkmal Geburtsort für die vorliegenden Daten nur eine geringe Varianz vorweist. Daher ist das Minimalszenario wahrscheinlich mit den Linkage-Ergebnissen vergleichbar, die beim Vorliegen eines Geburtsorts erreicht werden würden.

Obwohl die Ergebnisse der getesteten Szenarien für weitaus restriktivere Verfahren erstaunlich gut sind, werden diese Szenarien bei einem bundesweiten Register aufgrund der höheren Anzahl Records und der daraus resultierenden hohen Wahrscheinlichkeit für Verwechslungen mit allergrößter Wahrscheinlichkeit zu deutlich schlechteren Ergebnissen führen. Daher wird auf ein Verfahren mit einer noch höheren Datensparsamkeit verzichtet.

Aufgrund fehlender Merkmale im ZSR können die von Schnell (2019a: 4) vorgeschlagenen Szenarien nicht getestet werden. Die in der Expertise gezeigte Bedenklichkeit gegenüber der Verwendung von Adressdaten wird im mittleren Szenario berücksichtigt. Anstatt der Adresse wird die 100m-Gitter-ID verwendet. Alle weiteren Merkmale werden aus dem Minimalszenario übernommen. Dies entspricht dem Szenario 8, welches zudem den höchsten F^*-Wert aller Szenarien enthält.

Um dennoch die Wirksamkeit der Adresse zu testen, wird im Maximalszenario die vollständige Anschrift verwendet. Gegenüber dem Minimalszenario stellt das Maximalszenario somit den Fall mit der geringsten Datensparsamkeit dar. Das Maximalszenario entspricht dem Szenario 6.

Die gewählten drei Szenarien werden in Tabelle 5.3 dargestellt. Zur besseren Lesbarkeit der Ergebnisse werden die ausgewählten Szenarien im Folgenden als Szenario A, B und C bezeichnet. Das Szenario A ist das Minimalszenario, das Szenario C das Maximalszenario. Die Staatsbürgerschaft wird nicht als Merkmal verwendet. Wie oben beschrieben, ist das Merkmal irrelevant für die Verknüpfung des ZSR und damit auch eines bundesweiten Bildungsverlaufsregisters.

QID	Szenario A	Szenario B	Szenario C
Vorname	X	X	X
Nachname	X	X	X
Geschlecht	X	X	X
Geburtstag	X	X	X
Geburtsmonat	X	X	X
Geburtsjahr	X	X	X
100m-Gitter-ID		X	
Straße			X
Hausnummer			X
Postleitzahl			X

Tabelle 5.3.: Ausgewählte Linkage-Szenarien.

Tabelle 5.4 stellt die Linkage-Ergebnisse und den Linkage-Bias der ausgewählten drei Szenarien dar. Die Interpretation der Ergebnisse folgt den in Abschnitt 5.3 gegeben Interpretationen. Der beste F^*-Wert wird in Szenario B unter Verwendung des modifizierten probabilistischen Record-Linkage erreicht. Alle Verfahren, mit Ausnahme der Multiple Matchkeys, erreichen eine hohe Precision. Der exakte Abgleich verursacht in allen Szenarien keine false positives.[1]

Die Ergebnisse des exakten Matchkeys zeigen erneut die Wirksamkeit von Matchkeys zur Filterung von leicht zu verknüpfenden Paaren. Etwa 98 % der Paare können in Szenario A bereits durch den exakten Matchkey richtig zugeordnet werden. Eine einzelne Betrachtung der exakten Matchkeys ist jedoch nicht sinnvoll, da hierzu ein zu schlechter Recall vorliegt. In Form eines Filters sind exakte Matchkeys jedoch unverzichtbar.

Auf eine Darstellung der Ergebnisse für den Schwellenwert 0.9 der Multiple Matchkeys wird in Tabelle 5.4 aus Platzgründen verzichtet (die Zahlen finden sich in Tabelle A.1). Bereits der höchste und niedrigste Schwellenwert zeigen deutlich, dass in keinem Szenario eine gute Linkage-Qualität erzeugt wird ($\max(F^*) = 0.965$). Insbesondere die Zahl der false positives ist inakzeptabel hoch. Aus diesem Grund wird auf eine weitere Betrachtung der Methode verzichtet.

Sowohl beide Versionen des probabilistischen Record-Linkage als auch das Datavant-Verfahren und die CLKs liefern eine hohe Linkage-Qualität. Diese Verfahren werden daher miteinander verglichen. Für die CLKs und das probabilistischen Record-Linkage werden jeweils die Ergebnisse mit dem Schwellenwert 0.8 untersucht. Der jeweils höhere Schwellenwert wird nicht gewählt, da die zumeist geringe Reduktion von false positives mit einer zumeist deutlichen Erhöhung false negatives einherging.

Die Ergebnisse zum Linkage-Bias zeigen, dass – mit Ausnahme der Umzüge – kein Verfahren für die durchgeführte Verknüpfung verursacht. Unter Verwendung der vollständigen Adresse können insbesondere CLKs Personen nicht verknüpfen, die umgezogen sind. Dies liegt an einer zu hohen Gewichtung der Adresse. Das Vorgehen ist wie bereits beschrieben nicht ratsam und wird nur für eine konsistente Darstellung präsentiert. Bessere Ergebnisse könnten in diesem Fall durch das Auslassen der Adresse erreicht werden (Szenario A). Es muss beachtet werden, dass die durchgeführte Verknüpfung eine einmalige Verknüpfung über einen relativ kurzen Zeitraum darstellt. Dadurch verursachen auch extreme Record-Linkage-Methoden, wie das exakte Verfahren, keinen Linkage-Bias. Da zur Erstellung eines bundesweiten Bildungsverlaufsregisters mehr Verknüpfungen über einen deutlich längeren Zeitraum durchgeführt werden müssen, lassen sich diese Ergebnisse somit nicht direkt übertragen. Dies wird weiter in Abschnitt 5.4.4 untersucht.

[1] Etwaige false positives durch die betrachteten exakten Matchkeys waren auch nicht möglich, da alle false positives durch das Clerical-Review umcodiert wurden (Abschnitt 4.2).

Sz.	Methode	Schw.	n_k	TP	FP	FN	Prec.	Recall	F^*	d_{Alter}	$V_{Sf.}$	h_{Quelle}	$h_{Mig.}$	$h_{Mig.-h.}$	$h_{Deu.}$	um.(%)
A	Exact			178838	0	3547	**1.0000**	0.9806	0.9806	0.02	0.01	-0.02	-0.01	0.00	0.01	96.23
	MMK	0.70	2	179839	4276	2546	0.9768	0.9860	0.9635	-0.03	0.01	0.03	-0.01	0.00	0.01	97.36
		0.98	2	179839	4276	2546	0.9768	0.9860	0.9635	-0.03	0.01	0.03	-0.01	0.00	0.01	97.36
	CLK	0.80		180051	81	2334	0.9996	0.9872	0.9868	0.01	0.00	-0.01	-0.01	0.00	0.00	97.67
		0.85		179719	30	2666	0.9998	0.9854	0.9852	0.01	0.00	-0.01	-0.01	0.00	0.00	97.31
	DTV		4	180893	22	1492	0.9999	0.9918	0.9917	0.00	0.00	-0.01	-0.01	0.00	0.00	98.57
	ECM	0.80		180525	22	1860	0.9999	0.9898	0.9897	0.01	0.00	-0.01	-0.01	0.00	0.00	98.23
		0.90		180449	1	1936	1.0000	0.9894	0.9894	0.01	0.00	-0.01	-0.01	0.00	0.00	98.14
	ECM(D)	0.80		181074	45	1311	0.9998	**0.9928**	**0.9926**	0.00	0.00	-0.01	0.00	0.00	0.00	98.72
		0.90		181062	41	1323	0.9998	0.9927	0.9925	0.00	0.00	-0.01	0.00	0.00	0.00	98.72
B	Exact			163340	0	19045	**1.0000**	0.8956	0.8956	0.03	0.02	-0.04	-0.04	0.00	0.02	3.29
	MMK	0.70	4	180706	6632	1679	0.9646	0.9908	0.9560	-0.06	0.02	0.06	0.00	-0.01	0.01	93.89
		0.98	3	181214	5401	1171	0.9711	0.9936	0.9650	-0.05	0.02	0.05	-0.01	-0.01	0.01	96.31
	CLK	0.80		180758	166	1627	0.9991	0.9911	0.9902	0.01	0.00	0.00	0.00	0.00	0.00	97.98
		0.85		179963	32	2422	0.9998	0.9867	0.9865	0.01	0.00	-0.01	-0.01	0.00	0.00	96.64
	DTV		10	181759	57	626	0.9997	0.9966	0.9963	0.00	0.00	0.00	0.00	0.00	0.00	98.62
	ECM	0.80		181646	59	739	0.9997	0.9959	0.9956	0.00	0.00	0.00	-0.01	0.00	0.00	97.77
		0.90		181646	59	739	0.9997	0.9959	0.9956	0.00	0.00	0.00	-0.01	0.00	0.00	97.77
	ECM(D)	0.80		182005	92	380	0.9995	**0.9979**	**0.9974**	0.00	0.00	0.00	0.00	0.00	0.00	98.53
		0.90		182004	92	381	0.9995	0.9979	0.9974	0.00	0.00	0.00	0.00	0.00	0.00	98.53
C	Exact			162360	0	20025	**1.0000**	0.8902	0.8902	0.03	0.03	-0.05	-0.04	0.00	0.03	0.00
	MMK	0.70	4	167658	10065	14727	0.9434	0.9193	0.8712	-0.07	0.03	0.07	-0.03	-0.01	0.03	16.17
		0.98	7	169607	3753	12778	0.9784	0.9299	0.9112	-0.02	0.01	0.02	-0.03	0.00	0.02	26.09
	CLK	0.80		166364	65	16021	0.9996	0.9122	0.9118	0.01	0.01	-0.02	-0.03	0.00	0.02	8.29
		0.85		165677	17	16708	0.9999	0.9084	0.9083	0.02	0.02	-0.03	-0.03	0.00	0.02	7.71
	DTV		7	181683	60	702	0.9997	0.9962	0.9958	0.00	0.00	0.00	0.00	0.00	0.00	98.58
	ECM	0.80		181934	85	451	0.9995	**0.9975**	**0.9971**	0.00	0.00	0.00	0.00	0.00	0.00	97.62
		0.90		181930	85	455	0.9995	0.9975	0.9970	0.00	0.00	0.00	0.00	0.00	0.00	97.60
	ECM(D)	0.80		181740	338	645	0.9981	0.9965	0.9946	0.00	0.00	0.00	0.00	0.00	0.00	96.48
		0.90		170360	336	12025	0.9980	0.9341	0.9324	0.01	0.01	-0.02	-0.03	0.00	0.02	28.09

Tabelle 5.4.: Linkage-Ergebnisse und -Bias der ausgewählten Szenarien. Für Precision, Recall und F^* ist jeweils das Maximum innerhalb eines Szenarios dick gedruckt (Sz. = Szenario; Schw. = Schwellenwert; n_k = Anzahl verwendeter Matchkeys; Prec. = Precision; Sf. = Schulform; Quelle = Daten stammen von einer Schule; Mig. = Migrant; Mig.h. = Migrationshintergrund; Deu. = Deutsch; um. = Anteil richtig verknüpfter umgezogener Personen; d = Cohens's d; h = Cohen's h; V = Cramer's V; MMK = Multiple Matchkeys; DTV = Datavant; ECM(D) = ECM mit Dice-Modifikation).

		Precision			Recall		
Szenario	Verfahren	Melderegister	Schule	Beide	Melderegister	Schule	Beide
A	CLK	1.0000	0.9999	1.0000	0.9949	0.9560	0.9379
	Datavant	1.0000	1.0000	0.9999	0.9950	0.9731	0.9775
	ECM	1.0000	1.0000	0.9999	0.9946	0.9678	0.9610
	ECM(D)	1.0000	1.0000	1.0000	0.9958	0.9751	0.9789
B	CLK	1.0000	0.9999	0.9999	0.9958	0.9708	0.9613
	Datavant	1.0000	1.0000	0.9999	0.9981	0.9881	0.9889
	ECM	1.0000	1.0000	1.0000	0.9976	0.9874	0.9867
	ECM(D)	1.0000	0.9999	1.0000	0.9990	0.9906	0.9942
C	CLK	1.0000	0.9999	0.9998	0.9247	0.8412	0.8512
	Datavant	1.0000	1.0000	0.9996	0.9979	0.9882	0.9856
	ECM	1.0000	0.9999	0.9999	0.9986	0.9902	0.9933
	ECM(D)	0.9994	0.9997	0.9990	0.9979	0.9865	0.9910

Tabelle 5.5.: Precision und Recall nach Datenquelle für die ausgewählten Linkage-Szenarien und -Verfahren.

5.4.2. Unterschiede in der Linkage-Qualität zwischen Melderegister- und Schuldaten

Die Betrachtung der Datenqualität zeigte, dass insbesondere zwischen Melderegisterdaten und Schuldaten sowie zwischen den Herkunftsgruppen unterschiedliche Fehlerquoten vorhanden sind. Die unterschiedlichen Datenquellen stellen eine Besonderheit des ZSRs dar, welche die Interpretation der Linkage-Ergebnisse maßgeblich beeinflusst. So offenbarte die Darstellung der Datenflüsse des ZSR (Abschnitt 4.1.1), dass zunächst alle Personen durch die Schule gemeldet und nachträglich für alle schulpflichtigen Kinder die Daten aus dem Melderegister übernommen werden. Dies bedeutet, dass die aus dem Melderegister übernommenen Daten vorverarbeitet und mit den Schuldaten bereits einmal verknüpft wurden.

Werden für ein bundesweites Bildungsverlaufsregister rohe Schuldaten verwendet, müssen die Linkage-Ergebnisse des ZSR getrennt nach Datenquelle betrachtet werden. Tabelle 5.5 stellt hierzu Precision und Recall für die ausgewählten Szenarien und Verfahren nach Datenquelle dar. Für alle Verfahren zeigt sich, dass mit geringen Einbrüchen in der Precision zu rechnen ist. Deutlich stärkere Unterschiede zeigen sich beim Recall. Besonders in Szenario A wurden die Schuldaten bei allen Verfahren schlechter verknüpft als die Melderegisterdaten. Dies zeigt, dass für ein bundesweites Register auf Basis von rohen Schuldaten das Szenario A ungeeignet ist. Gute Ergebnisse können nur durch Verwendung eines Adressmerkmals erreicht werden. Die besten Ergebnisse liefert dabei das modifizierte probabilistische Record-Linkage unter Verwendung der 100m-Gitter-ID (Szenario B).

Da die Schulen besonders Namen häufiger unvollständig erfassen, muss zudem bei der Größe eines bundesweiten Registers davon ausgegangen werden, dass die Wahrscheinlichkeit für Verwechslungen steigt. Dementsprechend wird eine Verwendung von rohen Schuldaten wahrscheinlich zu einem noch niedrigeren Recall führen, als die Ergebnisse aufzeigen.

Für ein bundesweites Register müssen daher Maßnahmen getroffen werden, um eine hohe Qualität der gemeldeten Daten zu garantieren. Dies kann z. B. durch Erfassung des Meldeorts und einem Abgleich mit dem zugehörigen Einwohnermelderegister erfolgen. Nicht zuordenbare Fälle können so ggf. frühzeitig erkannt werden. Eine Sicherstellung einer hohen Datenqualität durch die Schulen selbst erscheint hingegen als keine tragfähige Lösung. So können besonders Probleme durch Unterschiede

zwischen den Datenqualitäten der Schulen entstehen. Falls dies mit der Auslastung der Schule korreliert, kann ein Bildungs-Bias entstehen. Die Hypothese kann mit den vorliegenden Daten jedoch nicht untersucht werden.

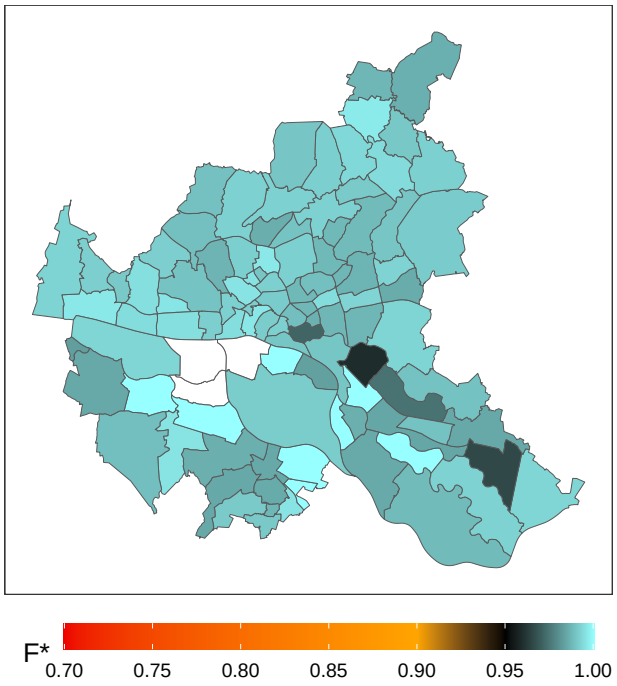
5.4.3. Geografische Unterschiede in der Linkage-Qualität

Für kohärente und vergleichbare Ergebnisse dürfen nur geringe Unterschiede in der Linkage-Qualität zwischen Regionen vorhanden sein. Dies wird im Folgenden für alle Personen, die im Schuljahr 2021/22 in Hamburg lebten und mit einem der ausgewählten Verfahren verknüpft wurden, untersucht. In den Abbildungen 5.3, 5.4 und 5.5 werden hierzu die F^*-Wert nach Hamburger Stadtteil dargestellt. Die im Zentrum erkennbaren weißen Stadtteile sind aufgrund des Hamburger Hafens unbewohnt.

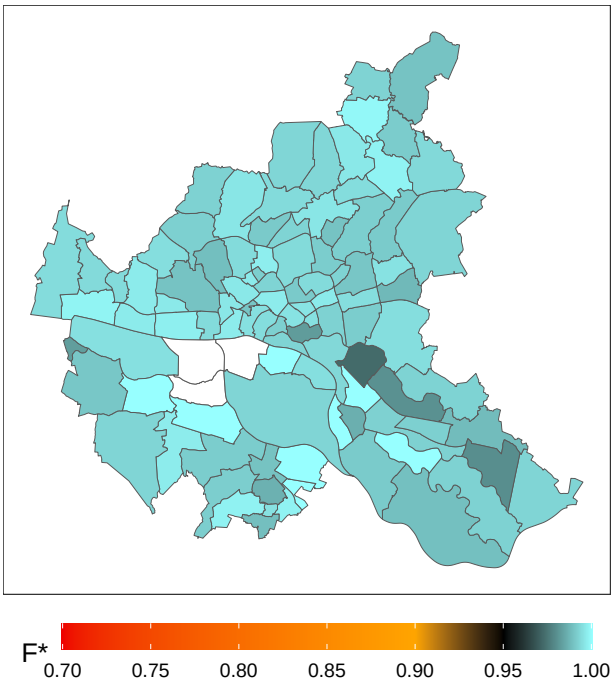
Bei allen Szenarien fallen insbesondere die Stadtteile Billwerder, Billbrook und Curslack auf. Die Lage der genannten Stadtteile wird in Abbildung 5.6 dargestellt. Auffällig für diese Stadtteile ist zunächst, dass sie zu den Stadtteilen mit einer geringeren Population in Hamburg gehören. So befinden sich im ZSR im Schuljahr 2021/22 614 Personen aus Curslack, 391 Personen aus Billbrook und 492 Personen aus Billwerder. Billbrook und Billwerder gehören zudem zu den Stadtteilen mit dem höchsten Anteil Menschen mit Migrationshintergrund in Hamburg. Im Jahr 2021 lag der Anteil für Billbrook bei 85.4 % und für Billwerder bei 61.6 %, in Curslack jedoch nur bei 28.5 % (Statistikamt Nord 2023).

Insbesondere für den Stadtteil Curslack können keine spezifischen Gründe für die häufig erkennbare niedrigere Linkage-Qualität gefunden werden. Möglich ist daher auch eine zufällige geografische Konzentration von schwer zu verknüpfenden Fällen. Gegeben geringen Anzahl Einwohner in Curslack würden bereits 14 falsch verknüpfte Paare ausreichen, um eine F^* von 0.97 zu erreichen.[1] Eine detaillierte Analyse ist aufgrund der geringen Anzahl Fehlverknüpfungen je Stadtteil nicht möglich. Die auffällig schlechte Linkage-Qualität der Stadtteile Billwerder und Billbrook sind mit großer Wahrscheinlichkeit nicht zufällig.

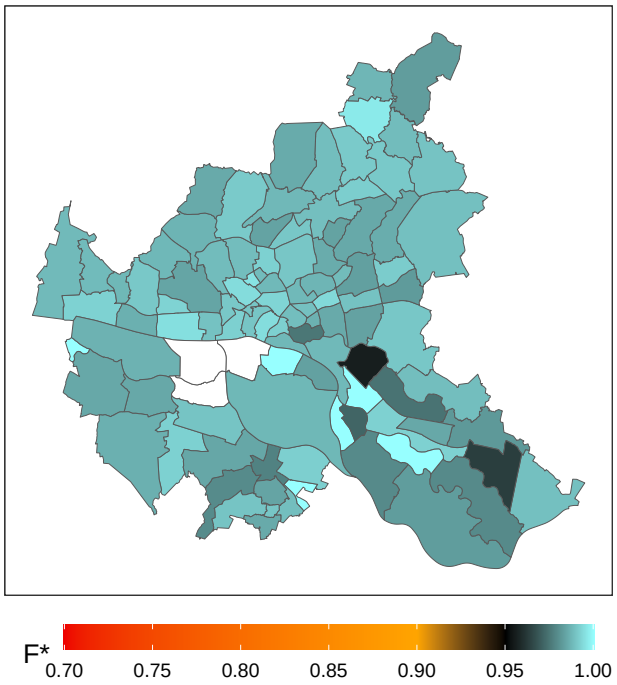
[1] Aus der Bevölkerung in Curslack können 488 Personen richtig verknüpft werden. Bei 14 false negatives ergibt sich hieraus $\frac{488-14}{488} = 0.971$. Bei 14 false positives ergibt sich $\frac{488}{488+14} = 0.972$.



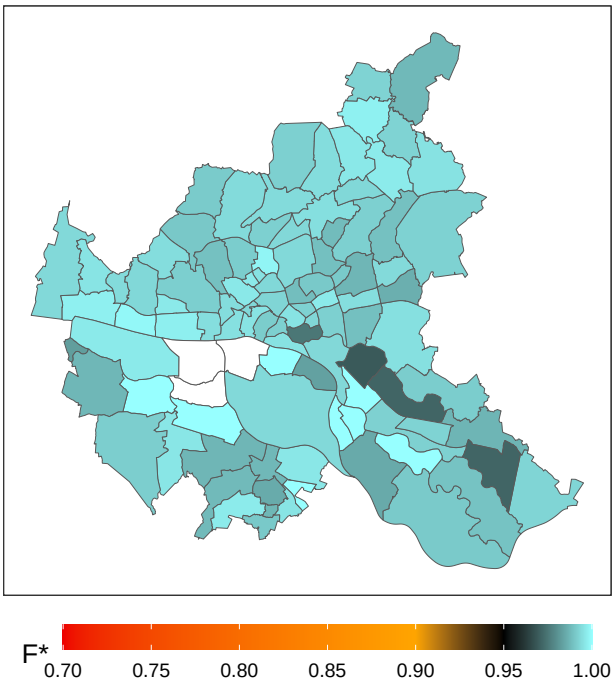
(a) Probabilistisches Record-Linkage



(b) Modifiziertes probabilistisches Record-Linkage

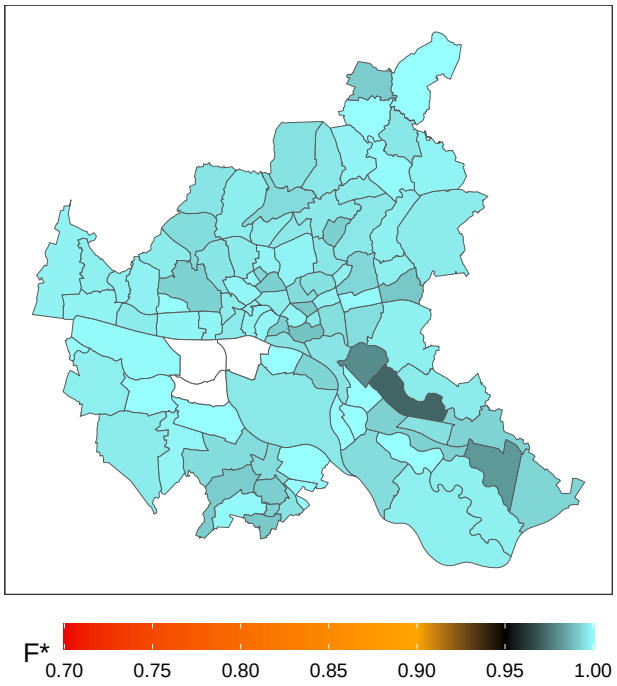


(c) CLKs

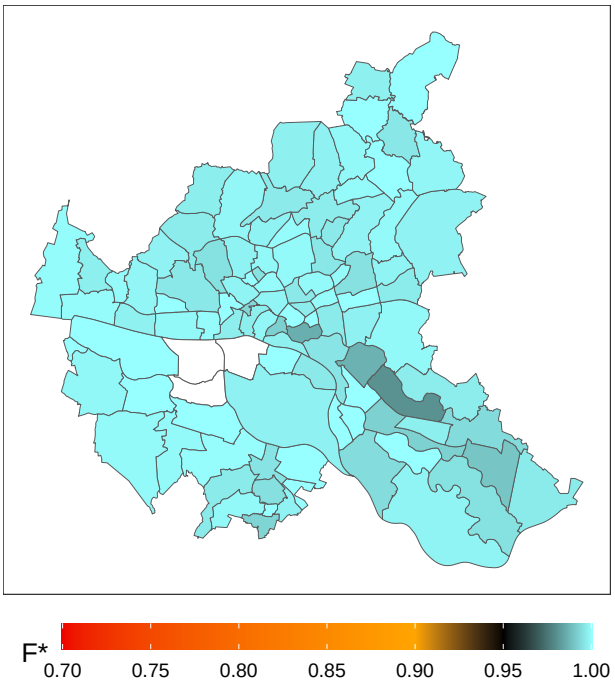


(d) Datavant

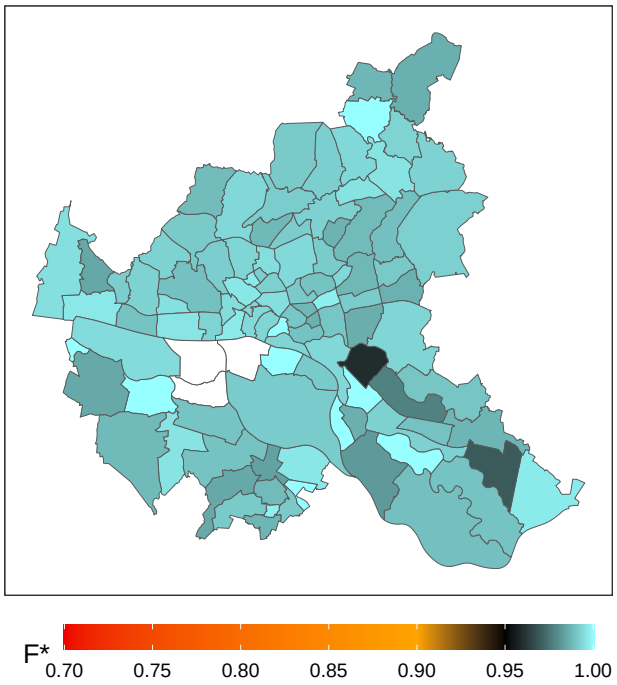
Abbildung 5.3.: Linkage-Qualität gruppiert nach Hamburger Stadtteil (ohne Neuwerk) für die ausgewählten Record-Linkage-Verfahren im Szenario A.



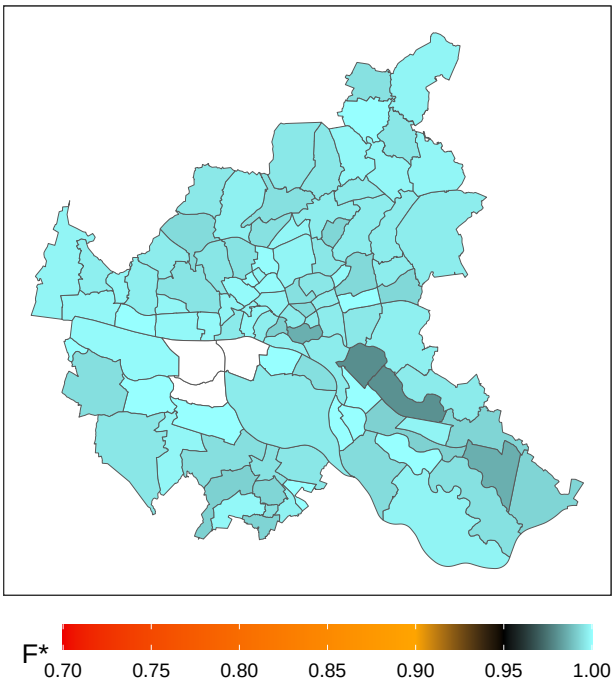
(a) Probabilistisches Record-Linkage



(b) Modifiziertes probabilistisches Record-Linkage

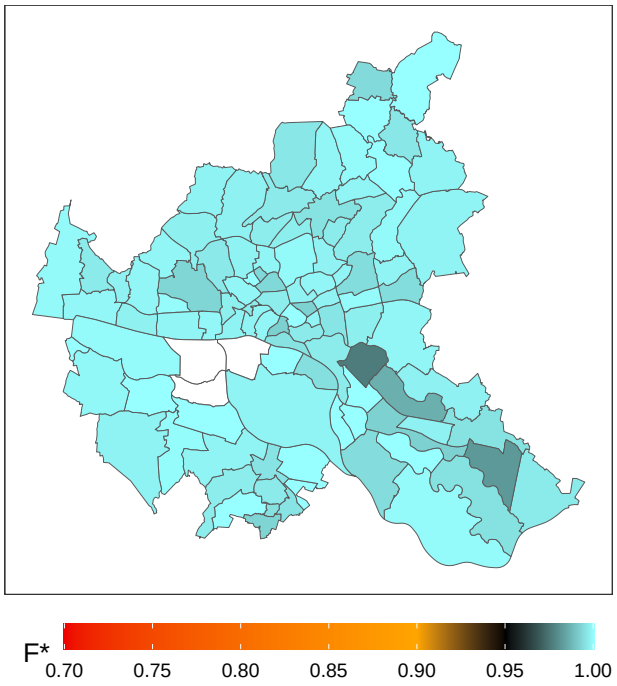


(c) CLKs

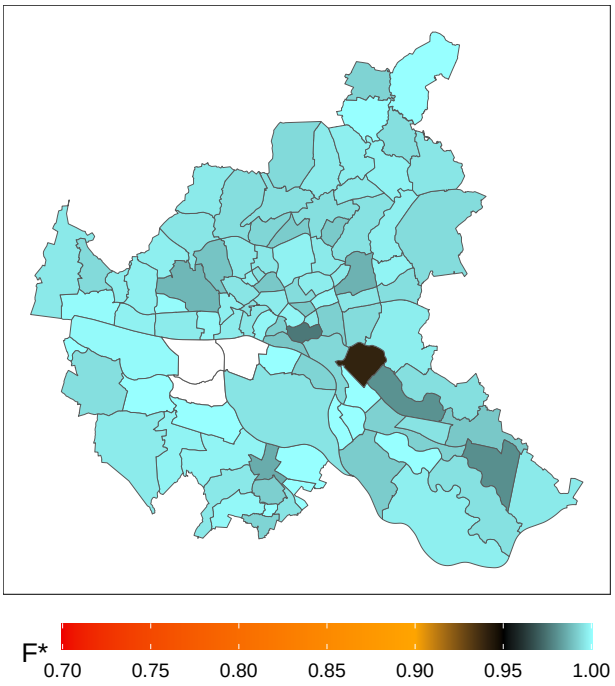


(d) Datavant

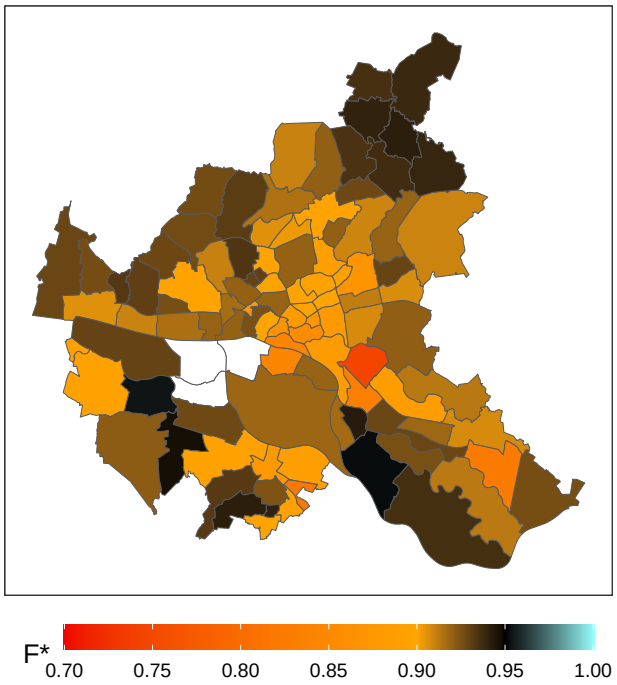
Abbildung 5.4.: Linkage-Qualität gruppiert nach Hamburger Stadtteil (ohne Neuwerk) für die ausgewählten Record-Linkage-Verfahren im Szenario B.



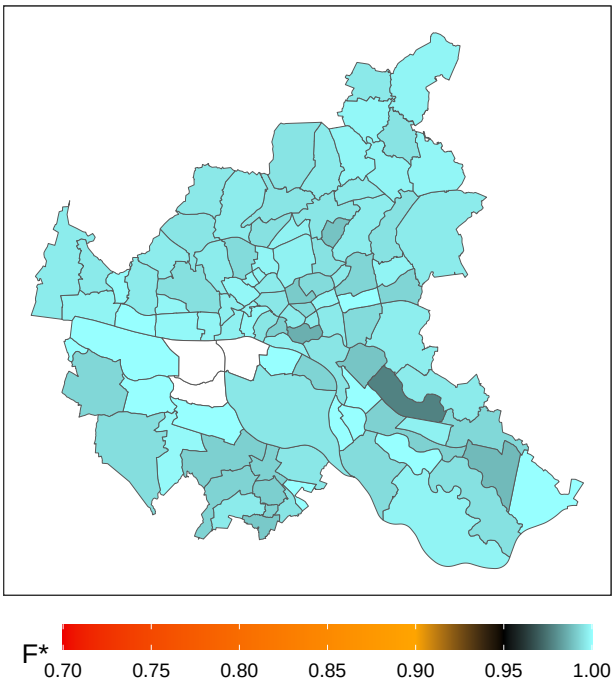
(a) Probabilistisches Record-Linkage



(b) Modifiziertes probabilistisches Record-Linkage



(c) CLKs



(d) Datavant

Abbildung 5.5.: Linkage-Qualität gruppiert nach Hamburger Stadtteil (ohne Neuwerk) für die ausgewählten Record-Linkage-Verfahren im Szenario C.

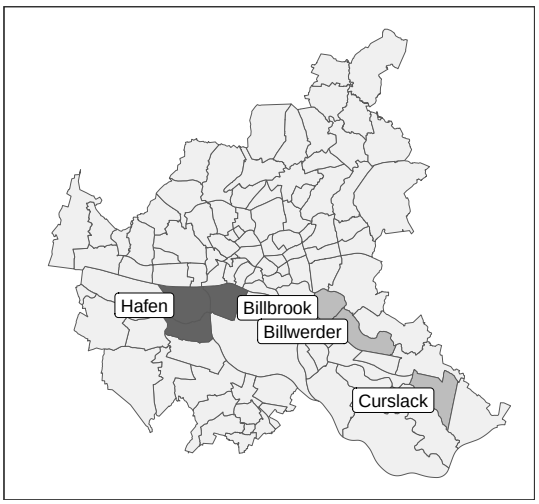


Abbildung 5.6.: Geografische Lage ausgewählter Hamburger Stadtteile und Gebiete.

Die Anforderung der Vergleichbarkeit und Kohärenz werden am besten durch das modifizierte probabilistische Record-Linkage in Szenario B sowie das Datavant-Verfahren in Szenario B und C erfüllt. Insgesamt zeigen die Ergebnisse, dass schwer zu verlinkende Fälle geografisch klumpen können.

Neben geografischen Unterschieden auf Stadtteilebene können Unterschiede auch kleinräumiger auftreten. In Abschnitt 4.3 und 4.5.2 zeigten sich bereits Auffälligkeiten bei Großgebäuden und Großwohnsiedlungen. Für das Record-Linkage können einige Probleme im Zusammenhang mit Großgebäuden auftreten. Erstens haben dort viele Menschen die gleiche Adresse und sind auf Basis der Adresse schwerer zu unterscheiden. Zweitens besteht die Möglichkeit, dass Menschen mit einem ähnlich soziokulturellen und demografischen Hintergrund geografisch deutlich enger klumpen, als es auf Stadtteilebene erkennbar ist. Zwar sind Großwohnsiedlungen unterschiedlich charakterisiert, jedoch bieten sie eine Vielzahl an Wohnungen, die zumeist ein ähnliches Bevölkerungssegment ansprechen (Kabisch 2021). Drittens können auch Flüchtlingsunterkünfte und -wohnheime zu Großgebäuden zählen bzw. werden Großgebäuden aufgrund des teilweise hohen Leerstands und der günstigen Wohnkosten als Flüchtlingsunterkünfte genutzt (Kabisch 2021).

Im Folgenden wird die Linkage-Qualität für Großgebäude und Großwohnsiedlungen anhand der Anzahl gemeldeter Schüler innerhalb einer 100m-Gitter-Zelle im Schuljahr 2021/22 betrachtet. Praktisch bedeutete dies, dass die Linkage-Qualität nach Bevölkerungsdichte untersucht wird. Dabei wird angenommen, dass die Bevölkerungsdichte stark mit Großgebäuden und Großwohnsiedlungen korreliert. Aufgrund der vorangestellten Betrachtung wird erwartet, dass die Linkage-Qualität mit zunehmender Bevölkerungsdichte abnimmt. Die Abbildungen 5.7, 5.8 und 5.9 zeigen die Linkage-Qualität gemessen am F^*-Wert nach Szenario und Record-Linkage-Verfahren.

In allen Szenarien zeigt sich, dass die Linkage-Qualität wie vermutet in dichter besiedelten Zellen abnimmt. Besonders interessant sind in diesem Zusammenhang die Ergebnisse des Szenarios A. Dieses Szenario enthält keine Adressinformationen und zeigte gleichzeitig die stärkste Verschlechterung der Linkage-Qualität bei zunehmender Anzahl Personen innerhalb der 100m-Gitterzelle. So beträgt die minimale Differenz zwischen der ersten und letzten Gruppe (1–20 und > 120) in Szenario A 0.02 (ECM modifiziert), in Szenario B 0.01 (ECM modifiziert) und in Szenario C 0.013 (ECM). Dies bedeutet, dass schwerer zu verknüpfende Personen geografisch in dichter besiedelten Gebieten klumpen. Der

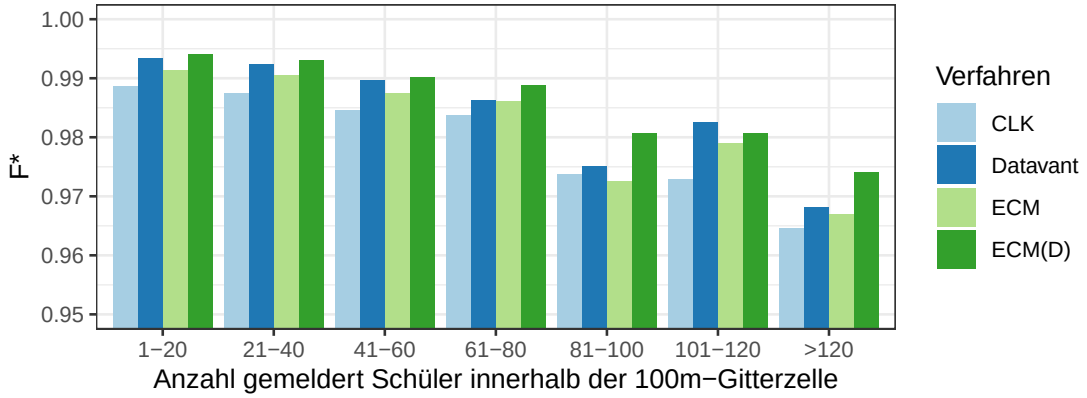


Abbildung 5.7.: Linkage-Qualität nach Bevölkerungsdichte und Record-Linkage-Verfahren für das Szenario A. Korrelationskoeffizienten der nicht gruppierten Daten: $r_{CLK} = -0.4$; $r_{Datavant} = -0.43$; $r_{ECM} = -0.37$; $r_{ECM(D)} = -0.36$.

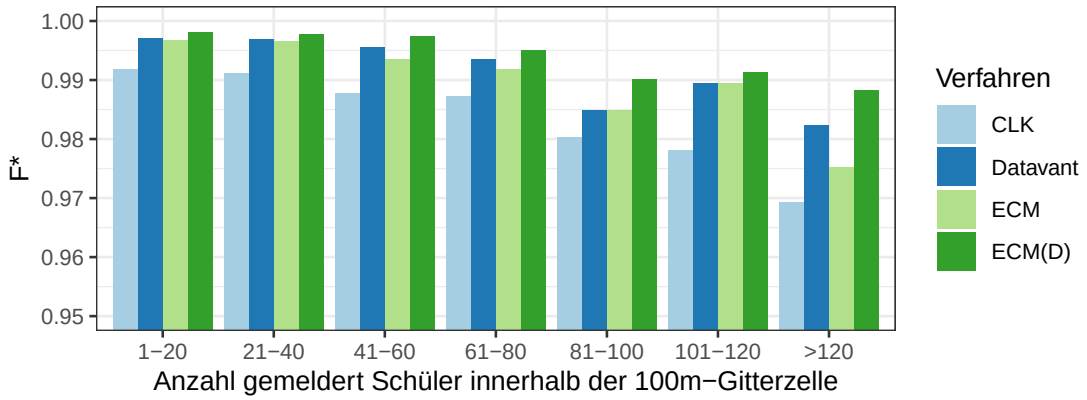


Abbildung 5.8.: Linkage-Qualität nach Bevölkerungsdichte und Record-Linkage-Verfahren für das Szenario B. Korrelationskoeffizienten der nicht gruppierten Daten: $r_{CLK} = -0.43$; $r_{Datavant} = -0.42$; $r_{ECM} = -0.44$; $r_{ECM(D)} = -0.37$.

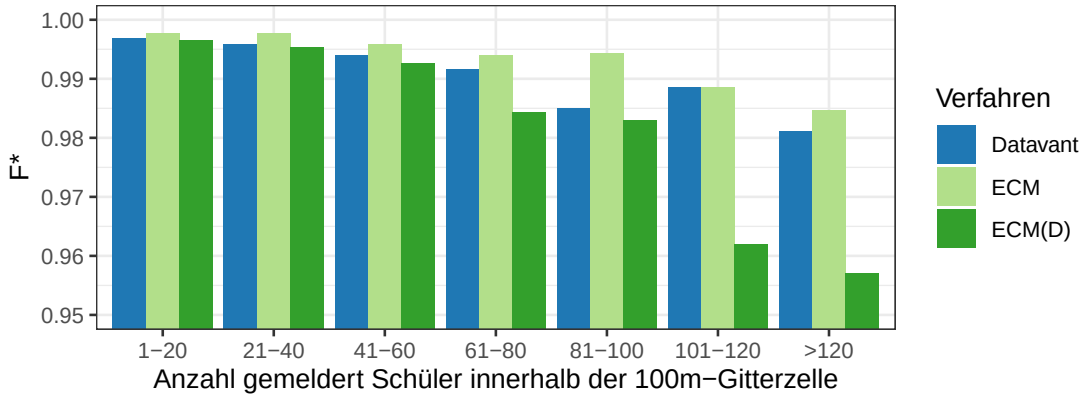


Abbildung 5.9.: Linkage-Qualität nach Bevölkerungsdichte und Record-Linkage-Verfahren für das Szenario C. Aufgrund der schlechten Linkage-Qualität werden die Ergebnisse der CLKs nicht dargestellt. Korrelationskoeffizienten der nicht gruppierten Daten: $r_{CLK} = -0.38$; $r_{Datavant} = -0.43$; $r_{ECM} = -0.4$; $r_{ECM(D)} = -0.46$.

Wohnort ist dabei keine Ursache für die schlechtere Linkage-Qualität, sondern steht nur in einem Zusammenhang.

Die Ergebnisse zeigen ferner, dass in Szenario B der Unterschiede bei der Linkage-Qualität etwas schwächer ist als in Szenario C. Insbesondere die modifizierte Variante des probabilistischen Record-Linkage weist zudem in Szenario C deutlich größere Probleme mit dicht besiedelten Gebieten auf. Die Ergebnisse deuten insgesamt darauf hin, dass die Verwendung der 100m-Gitter-ID die beste Vergleichbarkeit und Kohärenz produziert (Datenqualitätsdimension Abschnitt 3.3). Dennoch ist zu erwarten, dass in einem bundesweiten Bildungsverlaufsregister besonders Records von Personen, die in Großgebäuden wohnen, schlechter verknüpft werden.

5.4.4. Logistische Modelle zur Überprüfung eines Linkage-Bias

Die bisherigen Ergebnisse zeigen, dass kein Verfahren einen univariaten Linkage-Bias für soziodemografische Merkmale verursacht. Diese Ergebnisse lassen sich jedoch nicht direkt auf ein Bildungsverlaufsregister übertragen. So müssen für einen vollständigen Bildungsverlauf eine Vielzahl erfolgreicher Verknüpfungen durchgeführt werden. Dies führt dazu, dass bereits kleine systematische Unterschiede bei der Linkage-Qualität einen großen Unterschied beim Anteil vollständiger Bildungsverläufe verursachen kann. Wird z. B. eine Gruppe zu 99.9 % richtig verknüpft und der Rest der Population zu 100 %, so liegen nach 13 Schuljahren 1.3 % weniger vollständige Bildungsverläufe für die Gruppe im Vergleich zum Rest der Population vor.[1]

Die Einflüsse auf eine Fehlverknüpfung und auf einen daraus resultierenden unvollständigen oder falschen Bildungsverlauf müssen daher weiter untersucht werden. Hierzu wurde eine multivariate binärlogistische Regression zur Erklärung von Fehlverknüpfungen anhand soziodemografischer Merkmale durchgeführt. Die abhängige Variable des Modells ist die Gültigkeit der Verknüpfung. Alle richtig verknüpften Person sind mit dem Wert 1 codiert und alle falsch verknüpften Personen (false positive oder false negativ) mit dem Wert 0. Negative Koeffizienten im Modell sprechen somit für eine niedrigere Wahrscheinlichkeit, richtig verknüpft zu werden. Für alle Szenarien und Verfahren wurde ein einziges Modell entwickelt. Hierzu wurde berücksichtigt, dass zwar sehr viele Beobachtungen vorliegen, der Anteil falscher Verknüpfungen in einigen Szenarien jedoch sehr gering ist. Zur Vermeidung von Multikollinearität wurde sich auf eine geringe Zahl unabhängige Variablen beschränkt. Mithilfe des Varianzinflationsfaktors (VIF) wurde die Stärke der Multikollinearität geprüft.

Datenbasis waren die Merkmale der Personen aus dem Schuljahr 2021/22. Ein großes Problem stellten falsch verknüpfte Records dar, welche im Schuljahr 2023/24 nicht im ZSR enthalten sind. Da die fehlenden Informationen MNAR (Missing Not At Random) sind (Schnell 2019b: 162), konnten diese nicht imputiert werden. Daher wurden nur unabhängige Variablen verwendet, welche für ein Schuljahr vorliegen. Ferner sollen nur soziodemografische Faktoren untersucht werden und nicht der Einfluss von veränderten Ausprägungen, wie z. B. eine Änderung des Vornamens, auf die Verknüpfung.

Als unabhängige Variablen wurde der Migrationsstatus mit Deutsch als Referenzkategorie verwendet. Ein Interaktionseffekt zwischen dem Geschlecht und der Herkunft wurde getestet, führte jedoch zu keiner Modellverbesserung. Das Geschlecht wurde daher als eigenständige, unabhängige Variable dem Modell hinzugefügt. Um die grundsätzlich schlechtere Datenqualität von Schuldaten zu berücksichtigen, wurde die Datenquelle im Schuljahr 2021/22 als dichotome unabhängige Kontrollvariable hinzugefügt.

Die besuchte Schulform wurde dummy-codiert in das Model einbezogen. Da die Grundschule eine Einheitsschule ist, wurde sie als Referenzkategorie gewählt. Ein Problem stellte die geringe Anzahl

[1] Unter Annahme sonstiger Unabhängigkeit beträgt die Differenz der beiden Gruppen $1 - 0.999^{13} = 0.0129$.

Personen dar, die nicht eine Regelschule besuchen (Grundschule, Stadtteilschule, Gymnasium). Die übrigen Schulformen mussten daher zusammengefasst werden. Ein weiteres Problem bestand in den fehlenden Werten für Personen, die nicht im IVDS enthalten sind. Da nur die IVDS-Daten für allgemeinbildende Schulen vorliegen, wurde angenommen, dass alle Personen, die nicht im IVDS enthalten sind, Berufsschüler sind. Entsprechend wurde die Schulform Berufsschule ergänzt. Zur Einbeziehung des Wohnumfelds wurde der imputierte RISE-Status verwendet. Personen, die außerhalb Hamburgs lebten, wurde einer eigenen Kategorie zugewiesen. Aufgrund der geringen Anzahl Beobachtungen mit einem RISE-Status von 1 oder 2 wurde jeweils der Status 1-2 und 3-4 zusammengefasst. Nach der Datenaufbereitung lagen für alle beschriebene Merkmale keine fehlenden Werte vor.

Zur besseren Interpretation der Ergebnisse wurden die Average Marginal Effects (AMEs) berechnet (Hosmer et al. 2013: 317 ff.). Die Ergebnisse werden in Tabelle 5.6, 5.7, 5.8 und 5.9 für alle ausgewählten Verfahren und Szenarien dargestellt. Die entsprechenden Logits der Modelle finden sich im Anhang in den Tabellen A.2, A.3, A.4 und A.5.

Der Verknüpfungserfolg eines Record-Paars ist abhängig von der Ähnlichkeit der beiden Records und der Ähnlichkeit zu anderen vorliegenden Records. Die im Modell abgebildeten soziodemografischen Variablen stehen dadurch nur in einem Zusammenhang mit einer Fehlverknüpfung und können diese nicht vorhersagen. Dementsprechend ist es erwartbar, dass das Modell nur zu einem geringen Teil die Fehlverknüpfungen erklären kann. Dennoch liegt das Pseudo-R^2 bei vielen Verfahren bei etwa 0.1. Dies zeigt deutlich, dass soziodemografische Faktoren einen Einfluss auf die Wahrscheinlichkeit für eine richtige Verknüpfung haben.

Das Geschlecht hat nur selten einen signifikanten Effekt auf die Verknüpfungswahrscheinlichkeit. In den meisten Fällen ist der AME für Frauen geringfügig besser. Eine häufige Ursache für einen Geschlechtereffekt beim Record-Linkage besteht darin, dass Frauen häufiger einen Ehenamen annehmen als Männer. Da ein überwiegender Teil der erfassten Personen jünger als 18 Jahre ist, haben jedoch nur wenige Personen im Untersuchungszeitraum geheiratet. Daher ist ein Geschlechtereffekt durch Eheschließungen für die Daten auch nicht zu erwarten.

Bei allen Verfahren zeigt sich, dass der Herkunftseffekt am stärksten ist. So können Menschen mit Migrationshintergrund meist etwas schlechter als Deutsche verknüpft werden, Migranten meist deutlich schlechter. In den Verfahren mit einer besonders guten Linkage-Qualität zeigen sich nur kleine Unterschiede zwischen dem AME von Menschen mit Migrationshintergrund und Deutschen. Einzig Migranten können immer deutlich schlechter verknüpft werden. Der schwächste ermittelte Effekt beträgt -0.734 % und liegt in Szenario B beim modifizierten probabilistischen Record-Linkage vor. Die Effekte für Migranten unterscheiden sich zwischen dem modifizierten und dem unmodifizierten probabilistischen Record-Linkage zudem kaum. Entgegen der Erwartung löst die Verwendung der Blocknamen den Bias somit nicht weiter auf.

Bei der Schulform zeigt sich, dass die AMEs für Gymnasiasten zumeist genauso gut sind wie für Grundschüler, in manchen Fällen sogar etwas besser. Dem gegenüber wirkt sich der Besuch einer Stadtteil- oder Berufsschule signifikant schlecht auf den Verknüpfungserfolg aus. Besonders für Berufsschüler kann ein meist starker Effekt gefunden werden. Dabei besteht jedoch die Möglichkeit, dass der Effekt von Berufsschulen und der Datenquelle nicht vollständig getrennt werden konnte. So lag eine schwache Multikollinearität für den Faktor Berufsschule und Schuldaten vor (VIF > 2). Auffällig ist daher besonders der negative Effekt für Stadtteilschüler, welcher sich auch bei Verfahren mit einer guten Linkage-Qualität zeigt. Auch der Besuch anderer Schulform wirkte sich zumeist leicht negativ auf den Verknüpfungserfolg aus. Dies bedeutet, dass ein leichter Bildungs-Bias zugunsten Höherqualifizierter vorhanden ist. Die Ursache für diesen Bildungsbias ist jedoch unklar. Er kann sowohl durch eine andere Schülerschaft als auch durch eine andere Eingabepraxis verursacht werden.

Ein Effekt des RISE-Status und des Wohnorts kann nur bei einigen Verfahren beobachtet werden. Insbesondere bei Verwendung der CLKs im Szenario C zeigt sich ein starker Effekt für Personen, die außerhalb Hamburgs leben. Dieser Effekt ist darauf zurückzuführen, dass diese Personen häufig umgezogen sind und umgezogene Personen in diesem Szenario durch die CLKs schlecht verknüpft werden. Der starke Effekt ist somit auf das Verhalten dieser Personengruppe zurückzuführen. Die Datenquelle wirkt sich wie zu erwarten stark negativ auf den Verknüpfungserfolg aus.

Allgemein weist das Szenario A die stärksten und das Szenario B die schwächsten Effekte auf. Zwischen den Verfahren liegen nur geringe Unterschiede vor. Dabei unterscheiden sich die Richtungen der Effekte nur in wenigen Fällen. Die Effektstärken stehen zumeist in einem ähnlichen Verhältnis zueinander. Dies bedeutet, dass kein Verfahren einen spezifischen Linkage-Bias erzeugt. Die Verfahren unterscheiden sich nur in der Stärke des Linkage-Bias. Dabei liegt ein Linkage-Bias für alle Szenarien bei allen Verfahren vor. Für die Verknüpfung eines bundesweiten Bildungsverlaufsregisters bedeutet dies, dass ein Linkage-Bias bei den vollständig verknüpften Bildungsverläufe nur durch eine besonders gute Linkage-Qualität vermieden werden kann.

	Szenario A		Szenario B		Szenario C	
Variable	AME (%)	SE_{AME} (%)	AME (%)	SE_{AME} (%)	AME (%)	SE_{AME} (%)
Weiblich	0.083^{*}	0.047	0.053^{*}	0.031	0.030	0.025
Herkunft						
Deutsch						
Migrationshintergrund	-0.325^{***}	0.064	-0.177^{***}	0.038	-0.000	0.028
Migrant	-1.674^{***}	0.104	-1.529^{***}	0.090	-0.813^{***}	0.068
Schulform						
Grundschule						
Stadtteilschule	-0.412^{***}	0.059	-0.105^{***}	0.037	-0.099^{***}	0.030
Gymnasium	0.049	0.057	0.053	0.039	0.006	0.030
Berufsschule	-0.997^{***}	0.146	-0.424^{***}	0.094	-0.344^{***}	0.075
Andere	-0.497^{***}	0.165	-0.102	0.096	-0.144^{*}	0.084
Wohnort (RISE)						
Hoher Status (3-4)						
Niedriger Status (1-2)	-0.019	0.057	-0.002	0.035	0.041	0.028
Außerhalb Hamburgs	0.016	0.102	-0.201^{**}	0.089	-0.128^{**}	0.063
Schuldaten	-1.117^{***}	0.170	-0.349^{***}	0.093	-0.317^{***}	0.079
Log Pseudo-Likelihood	$-9,860.953$		$-4,578.595$		$-3,246.700$	
χ^2	1,229.640		1,104.931		721.911	
p	0.000		0.000		0.000	
R^2 (Nagelkerk)	0.062		0.110		0.102	
n	182,406		182,444		182,461	
FP+FN	1,881		798		527	

$^{*}p < 0,05$ $^{**}p < 0,01$ $^{***}p < 0,001$

Tabelle 5.6.: Average Marginal Effects (AME) und Standardfehler der AMEs (SE_{AME}) des logistischen Regressionsmodells für das unmodifizierte probabilistische Record-Linkage. Die abhängige Variable ist die Gültigkeit der Verknüpfung (0 = falsch verknüpft, 1 = richtig verknüpft).

	Szenario A		Szenario B		Szenario C	
Variable	AME (%)	SE_{AME} (%)	AME (%)	SE_{AME} (%)	AME (%)	SE_{AME} (%)
Weiblich	0.057	0.040	0.038	0.024	−0.001	0.032
Herkunft						
Deutsch						
Migrationshintergrund	−0.150***	0.052	−0.004	0.026	−0.076*	0.039
Migrant	−1.248***	0.089	−0.734***	0.063	−1.146***	0.078
Schulform						
Grundschule						
Stadtteilschule	−0.183***	0.052	−0.076***	0.027	−0.141***	0.038
Gymnasium	0.182***	0.050	0.019	0.027	0.066*	0.037
Berufsschule	−0.440***	0.111	−0.408***	0.082	−0.628***	0.103
Andere	−0.239*	0.138	−0.051	0.068	−0.281**	0.112
Wohnort (RISE)						
Hoher Status (3-4)						
Niedriger Status (1-2)	−0.061	0.049	0.012	0.027	−0.082**	0.040
Außerhalb Hamburgs	0.042	0.081	−0.082	0.054	−0.038	0.063
Schuldaten	−1.148***	0.173	−0.248***	0.068	−0.624***	0.109
Log Pseudo-Likelihood	−7,513.110		−2,909.144		−4,948.853	
χ^2	969.432		736.359		1,194.690	
p	0.000		0.000		0.000	
R^2 (Nagelkerk)	0.063		0.114		0.111	
n	182,430		182,476		182,615	
FP+FN	1,356		471		875	

*$p < 0,05$ **$p < 0,01$ ***$p < 0,001$

Tabelle 5.7.: Average Marginal Effects (AME) und Standardfehler der AMEs (SE_{AME}) des logistischen Regressionsmodells für das modifizierte probabilistische Record-Linkage. Die abhängige Variable ist die Gültigkeit der Verknüpfung (0 = falsch verknüpft, 1 = richtig verknüpft).

	Szenario A		Szenario B		Szenario C	
Variable	AME (%)	SE_{AME} (%)	AME (%)	SE_{AME} (%)	AME (%)	SE_{AME} (%)
Weiblich	0.157^{***}	0.053	0.066	0.046	-0.550^{***}	0.132
Herkunft						
Deutsch						
Migrationshintergrund	-0.223^{***}	0.072	-0.211^{***}	0.062	-1.344^{***}	0.172
Migrant	-1.450^{***}	0.104	-1.349^{***}	0.096	-8.545^{***}	0.264
Schulform						
Grundschule						
Stadtteilschule	-0.897^{***}	0.066	-0.509^{***}	0.056	0.715^{***}	0.166
Gymnasium	-0.229^{***}	0.061	-0.082	0.054	2.314^{***}	0.169
Berufsschule	-1.705^{***}	0.169	-1.241^{***}	0.147	-3.889^{***}	0.426
Andere	-0.986^{***}	0.191	-0.712^{***}	0.168	-0.965^{**}	0.467
Wohnort (RISE)						
Hoher Status (3-4)						
Niedriger Status (1-2)	-0.075	0.067	-0.057	0.056	-0.012	0.162
Außerhalb Hamburgs	0.290^{***}	0.094	-0.096	0.098	-2.722^{***}	0.447
Schuldaten	-1.700^{***}	0.199	-1.097^{***}	0.158	-2.614^{***}	0.405
Log Pseudo-Likelihood	$-11,999.946$		$-9,393.410$		$-52,885.997$	
χ^2	1,678.448		1,332.344		3,024.983	
p	0.000		0.000		0.000	
R^2 (Nagelkerk)	0.070		0.070		0.037	
n	182,465		182,548		182,441	
FP+FN	2,414		1,790		16,077	

$^{*}p < 0,05$ $^{**}p < 0,01$ $^{***}p < 0,001$

Tabelle 5.8.: Average Marginal Effects (AME) und Standardfehler der AMEs (SE_{AME}) des logistischen Regressionsmodells für das Record-Linkage mit CLKs und Multibit-Trees. Die abhängige Variable ist die Gültigkeit der Verknüpfung (0 = falsch verknüpft, 1 = richtig verknüpft).

	Szenario A		Szenario B		Szenario C	
Variable	AME (%)	SE_{AME} (%)	AME (%)	SE_{AME} (%)	AME (%)	SE_{AME} (%)
Weiblich	0.029	0.042	0.063^{**}	0.028	0.071^{**}	0.030
Herkunft						
Deutsch						
Migrationshintergrund	-0.319^{***}	0.057	-0.149^{***}	0.036	-0.232^{***}	0.039
Migrant	-1.555^{***}	0.097	-1.230^{***}	0.080	-1.278^{***}	0.082
Schulform						
Grundschule						
Stadtteilschule	-0.122^{**}	0.054	-0.101^{***}	0.036	-0.090^{**}	0.036
Gymnasium	0.260^{***}	0.052	0.083^{**}	0.036	0.068^{*}	0.038
Berufsschule	-0.540^{***}	0.125	-0.286^{***}	0.078	-0.375^{***}	0.092
Andere	-0.134	0.139	-0.039	0.084	-0.126	0.098
Wohnort (RISE)						
Hoher Status (3-4)						
Niedriger Status (1-2)	-0.064	0.051	-0.010	0.032	-0.039	0.035
Außerhalb Hamburgs	-0.040	0.094	-0.186^{**}	0.078	-0.010	0.069
Schuldaten	-1.089^{***}	0.169	-0.454^{***}	0.101	-0.418^{***}	0.101
Log Pseudo-Likelihood	$-8,188.397$		$-3,999.035$		$-4,431.854$	
χ^2	1,128.054		975.802		940.183	
p	0.000		0.000		0.000	
R^2 (Nagelkerk)	0.067		0.111		0.098	
n	182,405		182,440		182,439	
FP+FN	1,512		681		756	

$^{*}p < 0,05$ $^{**}p < 0,01$ $^{***}p < 0,001$

Tabelle 5.9.: Average Marginal Effects (AME) und Standardfehler der AMEs (SE_{AME}) des logistischen Regressionsmodells für das Record-Linkage nach dem Datavant-Verfahren. Die abhängige Variable ist die Gültigkeit der Verknüpfung (0 = falsch verknüpft, 1 = richtig verknüpft).

5.5. Übertragung der Linkage-Ergebnisse auf ein bundesweites Bildungsverlaufsregister

Alle bisherigen Ergebnisse zeigen, dass das Szenario A keine ausreichende Linkage-Qualität für ein Bildungsverlaufsregister erzeugt. Diese Ergebnisse werden im Folgenden in einen Zusammenhang mit den Ergebnissen der Simulation des Bildungsverlaufsregisters gebracht. Dadurch können die Ergebnisse der Verknüpfung des ZSR auf ein bundesweites Bildungsverlaufsregister übertragen werden.

Aufgrund der Größenunterschiede zwischen dem ZSR und einem bundesweiten Register können die in dieser Arbeit präsentierten Ergebnisse nicht direkt auf ein bundesweites Register übertragen werden. Daher bleiben die zentralen Ergebnisse der Simulation bestehen. So wird zur Verknüpfung eines bundesweiten Registers der Geburtsort als Merkmal benötigt. Wie bei der Auswahl der Szenarien beschrieben, kann das Szenario A mit dem Ergebnis der Simulation verglichen werden, da der Geburtsort im ZSR wahrscheinlich nur geringfügig variiert. Hieraus lässt sich zunächst ableiten, dass für ein bundesweites Register das Szenario A nicht ausreicht, um eine kohärente Statistik für große Gemeinden zu produzieren. Die Ergebnisse zeigen, dass zusätzlich ein Adressmerkmal benötigt wird, damit kein Linkage-Bias für Personen aus großen Gemeinden und dicht besiedelten Gebieten entsteht.

Ein möglicher Linkage-Bias nach Geburtsort wurde in den ursprünglichen Auswertungen der Simulation nicht untersucht. Dieser Linkage-Bias wurde daher nachträglich für die Simulationsergebnisse überprüft. Als Effektstärkemaß wurde Cramer's V verwendet. Für das letzte simulierte Jahr (2009) wurde der Zusammenhang zwischen der Gesamtverteilung der Geburtsorte und der Verteilung der Geburtsorte aller Personen berechnet, deren Bildungsverlauf vollständig verknüpft wurde. Für das Verfahren mit der besten Linkage-Qualität (probabilistisches Record-Linkage) und einer simulierten Fehlerquote von 0.3 % liegt ein Linkage-Bias von $V_{0.3\%} = 0.13$ vor (schwacher Effekt; Cohen 1988).[1] Auffällig sind wie erwartet große Gemeinden.[2]

Hieraus kann eindeutig geschlussfolgert werden, dass für ein bundesweites Bildungsverlaufsregister neben dem Geburtsort auch eine Adressangabe benötigt wird. Andernfalls können Personen aus großen Gemeinden schlechter verknüpft werden.

5.6. Linkage-Qualität für unterschiedliche Gittergrößen

Alle bisherigen Ergebnisse zeigen, dass die Verwendung der 100m-Gitter-ID zur besten Linkage-Qualität führt. Die Tests zur Einbeziehung einer größeren geografischen Distanz (Abschnitt 5.1.1) zeigen, dass mit inkonsistenten Ergebnissen schon bei einer geringen Distanz zu rechnen ist. Dabei muss beachtet werden, dass der berechnete Radius nicht zwangsläufig mit einem größeren Gitternetz übereinstimmt. Dies bedeutet, dass bei größeren Gitterzellen auch eine größere Fehlervarianz bei den Distanzergebnissen zu erwarten ist. Obwohl eine Verwendung eines 100m-Gitters in der amtlichen Statistik durchaus üblich ist, können datenschutzrechtliche Bedenken über die Genauigkeit der Adressangabe bestehen. Für eine endgültige Klärung der Veränderung der Datenqualität wird im folgenden Abschnitt die Linkage-Qualität für die 250m- und 500m-Gitterzellen betrachtet.

Um den Einfluss der unterschiedlichen Gittergrößen zu testen, wurden zunächst weitere Szenarien erstellt. Ähnlich wie bei den in Abschnitt 5.4 ausgewählten 3 Szenarien wurden die Merkmale Vorname,

[1] Ein entsprechend starker Bias zeigt sich auch bei allen anderen Fehlerquoten: $V_{0.1\%} = 0.14$, $V_{0.7\%} = 0.15$ $V_{1\%} = 0.16$

[2] Die Analyse der Simulationsergebnisse zeigt zudem, dass ein quadratischer Zusammenhang zwischen Anzahl Personen im Register und Anzahl unvollständiger Bildungsverläufe besteht ($r_{0.3\%} = 0.65$).

		Gitterzellengröße				
Gütemaß	Verfahren	100 m	250 m	500 m	1 km	PLZ
F*	CLK	0.9902	0.9901	0.9902	0.9901	0.9847
	ECM	0.9956	0.9956	0.9956	0.9956	0.9958
	ECM(D)	0.9974	0.9961	0.9961	0.9962	0.9962
Precision	CLK	0.9991	0.9986	0.9991	0.9992	0.9999
	ECM	0.9997	0.9997	0.9997	0.9997	0.9997
	ECM(D)	0.9995	0.9997	0.9997	0.9997	0.9997
Recall	CLK	0.9911	0.9914	0.9910	0.9909	0.9849
	ECM	0.9959	0.9960	0.9959	0.9960	0.9961
	ECM(D)	0.9979	0.9964	0.9964	0.9964	0.9965

Tabelle 5.10.: Linkage-Qualität der ausgewählten Record-Linkage-Verfahren bei Verwendung unterschiedlicher geografischer Gebiete.

Nachname, Geburtsdatum und Geschlecht beibehalten und darauf aufbauend die Gitter-ID variiert. Es wurden die gleichen Record-Linkage-Verfahren verwendet wie in Abschnitt 5.4. Aufgrund von fehlenden Matchkeys für die anderen Gittergrößen wurde jedoch das Datavant-Verfahren ausgeschlossen. Verwendet wurden daher nur CLKs und probabilistische Record-Linkage.

Die Linkage-Qualität der Szenarien wird in Tabelle 5.10 dargestellt. Zum Vergleich wird zusätzlich die Qualität bei Verwendung der Postleitzahl abgebildet (Szenario 7). Die CLKs weisen für alle Gittergrößen eine konstante Linkage-Qualität auf. Die Verwendung der Postleitzahl führt hingegen zu einer deutlichen Verschlechterung. Für das unmodifizierte probabilistische Record-Linkage können sehr geringe Unterschiede nach Gittergröße festgestellt werden. Dabei steigt die Linkage-Qualität bei größeren Gebieten geringfügig an. Das modifizierte probabilistischen Record-Linkage weist als einziges Verfahren eine deutliche Verbesserung des Recalls abhängig von der Gittergröße auf. So ist der Recall bei einem 100m-Gitter deutlich besser als bei allen anderen Gittergrößen. Dabei verschlechtert sich die Precision nur geringfügig.

Die Ergebnisse zeigen, dass eine deutliche Verbesserung der Linkage-Qualität nur bei Verwendung eines 100m-Gitters zu erwarten ist. Für größere Gitterzellen muss zudem das Kosten-Nutzen-Verhältnis betrachtet werden. Ein Nutzen bei der Verwendung von Gitterzellen besteht im Optimierungspotenzial für das Record-Linkage, durch die Berücksichtigung von geografischen Distanzen. Bei größeren Gitterzellen erhöht sich jedoch die Fehlervarianz dieser Distanz und das Optimierungspotenzial wird dadurch geringer. Da diese Optimierung im Rahmen dieser Arbeit nicht getestet werden kann, bleibt der endgültige Effekt dieser Fehlervarianz unklar. Ein weiterer Nutzen besteht darin, dass ein Gitternetz zur Analyse des Bildungsverlaufsregisters verwendet werden kann. Auch hierbei entstehen Probleme bei größeren Gitterzellen und der dadurch erhöhten Fehlervarianz. Die Ergebnisse bei Einbeziehung einer geografischen Distanz für das probabilistische Record-Linkage (Abbildung 5.2, S. 60) deuten darauf hin, dass entsprechende Probleme beim Record-Linkage bei einem 1km-Gitter auftreten werden. Dies bedeutet, dass ein 1km-Gitternetz einen geringen Mehrwert für die Linkage-Qualität und die Analyse der Daten besitzt. Ferner ist die Linkage-Qualität des probabilistischen Record-Linkage bei Verwendung der Postleitzahl ähnlich hoch wie bei größeren Gitterzellen. Die Erfassung der Postleitzahl ist zudem deutlich günstiger als die Erfassung der Gitter-ID, da die Postleitzahl erfragt werden kann, während die Gitter-ID durch ein System festgestellt werden muss. Dies bedeutet, dass sich die Kosten für die Implementierung und Pflege eines Systems zur Feststellung der Gitter-IDs nur für kleinere Gitterzellen rentieren. Idealerweise sollte ein 100m-Gitternetz verwendet werden. Für ein verschlüsseltes Record-Linkage mit CLKs wird die Verwendung eines Gitternetzes unbedingt empfohlen.

6. Zusammenfassung

Das Ziel dieser Arbeit bestand darin, weiteren Aufschluss über die Ausgestaltung eines bundesweiten Bildungsverlaufsregisters, welches über Record-Linkage erstellt werden soll, zu geben. Des Weiteren sollte die Belastbarkeit der Ergebnisse der durch Schnell (2022) durchgeführten Simulation des Registers geprüft werden. Dies erfolgte anhand von empirischen Daten des Hamburger Schülerregisters. Die durchgeführte Datenqualitätsanalyse von Quasi-Identifikatoren für Record-Linkage ist dabei einzigartig und mit keiner bisherigen Publikation vergleichbar.

Es muss hervorgehoben werden, dass Hamburg die zweitgrößte Gemeinde in Deutschland ist und größtenteils eine hohe Bevölkerungsdichte aufweist. Erstens ist dadurch das ZSR keine zufällige Auswahl. Zweitens kann erwartet werden, dass eine Gemeinde wie Hamburg die größten Probleme für das Record-Linkage eines bundesweiten Bildungsverlaufsregisters hervorrufen wird. So ist die Wahrscheinlichkeit für gleiche Merkmalsausprägungen (Wohnort und Geburtsort) und somit die Verwechslungswahrscheinlichkeit deutlich höher als bei kleineren Gemeinden. Dies zeigte sich auch bei einer nachträglichen Analyse der Simulationsergebnisse (Abschnitt 5.5). Daher stellen die präsentierten Ergebnisse einen Extremfall dar. Im Vergleich zu einem bundesweiten Bildungsverlaufsregister befinden sich im ZSR jedoch deutlich weniger Beobachtungen. Aus diesem Grund bleiben die grundlegenden Ergebnisse der Simulation des bundesweiten Registers bestehen. Eine ausreichende Linkage-Qualität für ein bundesweites Bildungsverlaufsregister ist nur bei einer Erfassung des Geburtsorts möglich.

Die Datenqualitätsanalyse des ZSR und IVDS zeigt, dass die Daten von Migranten eine höhere Fehlerquote aufweisen. Zudem ziehen Migranten häufiger und über weitere Distanzen um. Hieraus resultieren Probleme für das Record-Linkage, die sich auch in den Linkage-Ergebnissen zeigen. Obwohl in keinem Szenario ein univariater Linkage-Bias in Form eines schwachen Effektstärkemaßes ermittelt werden konnte, zeigen die Ergebnisse der logistischen Regressionen, dass in allen ausgewählten Verfahren ein Einfluss der Herkunft auf die Verknüpfungswahrscheinlichkeit besteht, sodass besonders Migranten schlechter verknüpft werden als Deutsche und Menschen mit Migrationshintergrund. Ferner konnte in allen ausgewählten Verfahren auch ein Einfluss der Schulform auf den Verknüpfungserfolg beobachtet werden, sodass besonders Schüler von Stadtteilschulen schlechter verknüpft werden als Schüler von Gymnasien und Grundschulen. Zudem bestehen auch geografische Unterschiede in der Linkage-Qualität. So klumpen schwer zu verknüpfende Fällen sowohl auf Stadtteilebene als auch in Großgebäuden. Die Linkage-Ergebnisse zeigen dabei, dass der Effekt besonders stark ist, wenn keine Adressdaten vorliegen.

Die genannten Effekte treten bei allen Record-Linkage-Verfahren auf. Dies bedeutet, dass ein im Zeitverlauf entstehender Linkage-Bias nur durch eine möglichst gute Verknüpfung aufgelöst werden kann. Die beste Linkage-Qualität konnte durch probabilistisches Record-Linkage erreicht werden. Dies war auch das Ergebnis der Simulation (Schnell 2022; Weiand 2022). Ferner zeigt die Datenqualitätsanalyse einige Optimierungsmöglichkeiten des probabilistischen Record-Linkage auf, die für eine Verbesserung der Linkage-Qualität genutzt werden können. Aufgrund des häufigen Fehlens von Token konnten die Ergebnisse durch die Verwendung von Blocknamen und die Berechnung des Dice-Koeffizienten zwischen den Mengen verwendeter Token zumeist verbessert werden. Weiteres Optimierungspotenzial besteht bei der Einbeziehung der geografischen Distanz.

Bei der Implementation dieser Optimierungen zeigte sich zudem, dass ein Mangel an geeigneter Record-Linkage-Software besteht. Aktuelle Record-Linkage-Software ist entweder schwierig zu erweitern, implementiert keine aktuellen Algorithmen oder skaliert nicht gut bei großen Datenmengen. Für die Implementation eines Verfahrens für ein bundesweites Register bedarf es somit genügend Ressourcen und Zeit.

Selbst bei der Implementierung aller vorgeschlagenen Optimierungen kann erwartet werden, dass nicht alle Personen automatisch richtig verknüpft werden können. Entsprechend benötigt es einer manuellen Prüfung von Fällen. Besondere Aufmerksamkeit sollte somit auch auf eine effiziente Filterung von zu prüfenden Fällen gelegt werden. Eine Möglichkeit hierfür besteht in der Kombination von Methoden mit einer hohen Linkage-Qualität.[1] Dies wurde durch die Kombination des Datavant-Verfahrens und des probabilistischen Record-Linkage gezeigt. Überdies führen falsch positive Verknüpfungen wahrscheinlich zu auffälligen Brüchen im Bildungsverlauf. Dies kann anhand der vorliegenden Daten nicht getestet werden, da nur die Daten für zwei Schuljahre vorliegen. Eine manuelle Prüfung von Fällen mit auffälligen Brüchen im Bildungsverlauf ist dennoch zu empfehlen. Entsprechende Regeln können aus bereits vorhanden empirischen Wissen zu Bildungsverläufen sowie juristischen und formalen Einschränkungen abgeleitet werden. Beispielsweise werden für ein Studium in der Regel bestimmte formale Voraussetzungen benötigt, die in einem Bildungsverlauf gegeben sein müssen. Sind diese Voraussetzungen nicht erfüllt, so sollte der entsprechende Fall manuell geprüft werden.

Aus der Notwendigkeit einer manuellen Klassifikation folgt zudem, dass ein bundesweites Bildungsverlaufsregister nicht durch eine vollständige Verschlüsselung erstellt werden kann. Die Linkage-Qualität der CLKs und des Datavant-Verfahrens sind hierfür nicht ausreichend. Sofern Zweifelsfälle nicht manuell in Klartext überprüft werden können, wird bei einem bundesweiten Bildungsverlaufsregister daher ein Linkage-Bias entstehen. Sollte eine solche Prüfung möglich sein, bedarf es darüber hinaus einiger methodischer Forschung zur Gewichtung von Merkmalen in CLKs. Die Gewichtung der QIDs stellt ein zentrales Optimierungspotenzial der CLKs dar.

Alle Ergebnisse zeigen, dass ein Adressmerkmal zur Verknüpfung eines bundesweiten Bildungsverlaufsregisters benötigt wird. Andernfalls entsteht ein Linkage-Bias für große Gemeinden wie Hamburg. Als Adressmerkmal wird jedoch keine vollständige Anschrift benötigt. Stattdessen eignet sich eine Gitterzellen-Koordinate nach dem INSPIRE-Referenzsystem, welche datensparsamer ist und bei der zahlreiche Vorteile für spätere Analysen erhalten bleiben. Die Ergebnisse des Record-Linkage zeigen, dass hierfür idealerweise ein 100m-Gitter benötigt wird.

Des Weiteren weisen die Linkage-Ergebnisse auf, dass die Staatsangehörigkeit ein irrelevantes Merkmal für das Record-Linkage eines bundesweiten Bildungsverlaufsregisters ist. Die Staatsangehörigkeit wird somit für das Record-Linkage nicht benötigt.

Die Linkage-Ergebnisse zeigen des Weiteren, dass besonders umgezogene Menschen schlechter verknüpft werden. Darüber hinaus stellen sich Namensänderungen auch bei jungen Menschen als ein Problem für die Verknüpfung heraus. Große Probleme bei einer Verknüpfung sind zu erwarten, wenn beide Fälle gleichzeitig auftreten. Dies ist besonders bei Unterbrechungen der Bildungslaufbahn wahrscheinlich.

Da ein erwarteter Nutzen des Registers die Analyse von Personen mit einer Unterbrechung in der Bildungslaufbahn ist, müssen weitere Maßnahmen eingerichtet werden, falls auch diese Fälle richtig verknüpft werden sollen. Es kann nicht erwartet werden, dass hierzu eine manuelle Prüfung ausreicht. Eine zuverlässige Klassifikation von schwierigen Fällen kann wahrscheinlich durch die Einrichtung und

[1] Zu prüfende Record-Paare können auch über die Match-Wahrscheinlichkeit anhand eines zweiten Schwellenwertes bestimmt werden. Oberhalb des ersten Schwellenwertes werden alle Paare als Match klassifiziert. Paare zwischen dem ersten und zweiten Schwellenwert werden manuell klassifiziert (Fellegi & Sunter 1969).

Verwendung weiterer Register gewährleistet werden. Anhand dieser Register können Veränderungen nachvollzogen und damit validiert werden. Damit Namensänderungen nachvollzogen werden können, wird ein Namensänderungsregister benötigt.[1] Zur Aufklärung von Adressänderungen benötigt es ein historisiertes Anschriftenregister.[2] Auf Grundlage des ursprünglichen Erfassungsdatums im bundesweiten Bildungsverlaufsregister können mithilfe dieser beiden Register mögliche alternative Werte zu einem gegebenen Zeitpunkt angegeben werden. Diese alternativen Werte können sowohl in das Record-Linkage als auch die manuelle Prüfung einbezogen werden, sodass höchstwahrscheinlich auch für Personen mit einer Unterbrechung im Bildungsverlauf kein Linkage-Bias entsteht. Im Zuge des geplanten Registerzensus gilt es zudem zu prüfen, inwieweit die vorgestellten Register bereits im Rahmen des Registerzensus benötigt und implementiert werden.

Die Datenqualitätsanalyse des ZSR zeigt, dass deutliche Unterschiede in der Datenqualität zwischen den Daten der Schulen und des Melderegisters bestehen. Angesichts dieser Unterschiede würde ein bundesweites Register auf Basis von Schuldaten höchstwahrscheinlich mit einem hohen Linkage-Bias einhergehen. Um einen möglichst hohen Verknüpfungserfolg zu erzielen, müssen die Schuldaten, wie im ZSR gehandhabt, mit einem Melderegister verknüpft und die Meldedaten an die Register der Länder weitergegeben werden. Sowohl die Praxis und der Erfolg des ZSR als auch die in dieser Arbeit präsentierten Record-Linkage-Ergebnisse weisen eindeutig darauf hin, dass ein unverzerrtes bundesweites Bildungsverlaufsregister, welches durch Record-Linkage erzeugt wird, möglich ist. Die in dieser Arbeit dargelegten Analysen und Informationen über das ZSR bieten hierfür eine wichtige Datengrundlage.

[1] Ein Namensänderungsregister muss mindestens den alten und neuen vollständigen Namen, das Geburtsdatum und das Datum der Änderung umfassen.

[2] Bei Verwendung eines Gittersystems muss ein historisiertes Anschriftenregister mindestens die alte und neue Gitter-ID sowie das Datum der Änderung umfassen.

Literatur

Baas, M. (2021). Bildungsbeteiligung nach Migrationshintergrund: Der Einfluss von Zuwanderungsgeneration, Zuzugsalter und Zuzugsmotiven. In: *WISTA - Wirtschaft und Statistik* 73.2, S. 111–125.

Batini, C. & M. Scannapieco (2006). *Data Quality: Concepts, Methodologies and Techniques*. Data-Centric Systems and Applications. Berlin: Springer.

Behörde für Schule und Berufsbildung (2021). Dienstanweisung Zentrales Schülerregister (DA ZSR). In: *Mitteilungsblatt der Behörde für Schule und Berufsbildung* 6, S. 87–90.

Behörde für Schule und Berufsbildung (2022). *Hamburger Schulstatistik: Schuljahr 2021/22*. Hamburg: Behörde für Schule und Berufsbildung.

Behörde für Schule und Berufsbildung (2024). *Hamburger Schulstatistik: Schuljahr 2023/24*. Hamburg: Behörde für Schule und Berufsbildung.

Behörde für Stadtentwicklung und Wohnen (2021). *Sozialmonitoring Integrierte Stadtteilentwicklung – Bericht 2021*. Hamburg: Behörde für Stadtentwicklung und Wohnen.

Behörde für Stadtentwicklung und Wohnen (2022). *Rahmenprogramm Integrierte Stadtteilentwicklung: Leitfaden für die Praxis*. Hamburg: Behörde für Stadtentwicklung und Wohnen.

Beicht, U. (2015). *Berufliche Orientierung junger Menschen mit Migrationshintergrund und ihre Erfolgschancen beim Übergang in betriebliche Berufsausbildung: Überblick über Ergebnisse quantitativer Forschung der letzten zehn Jahre in Deutschland sowie vergleichende Analysen auf Basis der BIBB-Übergangsstudien und der BA/BIBB-Bewerberbefragungen*. Wissenschaftliche Diskussionspapiere 163. Bonn: Bundesinstitut für Berufsbildung.

Bentley, J. P., J. B. Ford, L. K. Taylor, K. A. Irvine & C. L. Roberts (2012). Investigating Linkage Rates among Probabilistically Linked Birth and Hospitalization Records. In: *BMC Medical Research Methodology* 12.149.

Berk, R. A., B. Western & R. E. Weiss (1995). Statistical Inference for Apparent Populations. In: *Sociological Methodology* 25, S. 421–458.

Bernstam, E. V., R. J. Applegate, A. Yu, D. Chaudhari, T. Liu, A. Coda & J. Leshin (2022). Real-World Matching Performance of Deidentified Record-Linking Tokens. In: *Applied Clinical Informatics* 13.4, S. 865–873.

Bohensky, M. (2016). Bias in Data Linkage Studies. In: *Methodological Developments in Data Linkage*. Hrsg. von K. Harron, H. Goldstein & C. Dibben. Chichester: Wiley, S. 63–82.

Bohensky, M., D. Jolley, V. Sundararajan, D. V. Pilcher, S. Evans & C. A. Brand (2011). Empirical Aspects of Linking Intensive Care Registry Data to Hospital Discharge Data without the Use of Direct Patient Identifiers. In: *Anaesthesia and Intensive Care* 39.2, S. 202–208.

Bopp, M., J. Braun, D. Faeh & F. Gutzwiller (2010). Establishing a Follow-up of the Swiss MONICA Participants (1984-1993): Record Linkage with Census and Mortality Data. In: *BMC Public Health* 10.562.

Borgs, C. (2019). Optimal Parameter Choice for Bloom Filter-Based Privacy-Preserving Record Linkage. Diss. Universität Duisburg-Essen.

Bovee, M., R. P. Srivastava & B. Mak (2003). A Conceptual Framework and Belief-Function Approach to Assessing Overall Information Quality. In: *International Journal of Intelligent Systems* 18.1, S. 51–74.

Brameshuber, I. (2007). Auf dem Weg zu einem bundesweiten Schülerregister? In: *Recht der Jugend und des Bildungswesens* 55.3, S. 366–373.

Brown, A. P., C. Borgs, S. M. Randall & R. Schnell (2017). Evaluating Privacy-Preserving Record Linkage Using Cryptographic Long-Term Keys and Multibit Trees on Large Medical Datasets. In: *BMC Medical Informatics and Decision Making* 17.83.

Bundesministerium für Justiz (2023). *Die geplante Reform des Namensrechts: Häufig gestellte Fragen*. URL: https://www.bmj.de/SharedDocs/Downloads/DE/Themen/FamilieUndPartnerschaft/FAQ_Namensrecht.pdf?__blob=publicationFile&v=1 (zuletzt abgerufen am 23. 01. 2024).

Campbell, K. M. (2009). Impact of Record-Linkage Methodology on Performance Indicators and Multivariate Relationships. In: *Journal of Substance Abuse Treatment* 36.1, S. 110–117.

Chipperfield, J. O. & R. L. Chambers (2015). Using the Bootstrap to Account for Linkage Errors When Analysing Probabilistically Linked Categorical Data. In: *Journal of Official Statistics* 31.3, S. 397–414.

Christen, P. (2012). *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Berlin: Springer.

Christen, P. & A. Pudjijono (2009). Accurate Synthetic Generation of Realistic Personal Information. In: *Advances in Knowledge Discovery and Data Mining*. Hrsg. von T. Theeramunkong, B. Kijsirikul, N. Cercone & T.-B. Ho. Bd. 5476. Lecture Notes in Computer Science. Berlin: Springer, S. 507–514.

Christen, P., T. Ranbaduge & R. Schnell (2020). *Linking Sensitive Data: Methods and Techniques for Practical Privacy-Preserving Information Sharing*. Cham: Springer.

Christen, P. & R. Schnell (2023). Thirty-Three Myths and Misconceptions about Population Data: From Data Capture and Processing to Linkage. In: *International Journal of Population Data Science* 8.1.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. 2. Aufl. Hillsdale: Lawrence Erlbaum Associates.

Damerau, F. J. (1964). A Technique for Computer Detection and Correction of Spelling Errors. In: *Communications of the ACM* 7.3, S. 171–176.

Datavant (2020). *Overview of Datavant's De-Identification and Linking Technology for Structured Data*. URL: https://datavant.com/wp-content/uploads/2021/08/White-Paper_-De-Identifying-and-Linking-Structured-Data_updated-3-10-2020.pdf (zuletzt abgerufen am 06. 10. 2023).

De Bruin, J. (2015). Probabilistic Record Linkage with the Fellegi and Sunter Framework: Using Probabilistic Record Linkage to Link Privacy Preserved Police and Hospital Road Accident Records. Magisterarb. Delft University of Technology.

Deutsche Post (2024). *Postleitdaten*. URL: https://www.deutschepost.de/de/d/deutsche-post-direkt/datafactory/download_postleitdaten.html (zuletzt abgerufen am 14. 02. 2024).

Doidge, J. C. & K. L. Harron (2019). Reflections on Modern Methods: Linkage Error Bias. In: *International Journal of Epidemiology* 48.6, S. 2050–2060.

DuVall, S. L., A. M. Fraser, R. A. Kerber, G. P. Mineau & A. Thomas (2010). The Impact of a Growing Minority Population on Identification of Duplicate Records in an Enterprise Data Warehouse. In: *MEDINFO 2010: Proceedings of the 13th World Congress on Medical*

Informatics. Hrsg. von C. Safran, S. Reti & H. F. Marin. Amsterdam: IOS Press, S. 1122–1126.

Enamorado, T., B. Fifield & K. Imai (2019). Using a Probabilistic Model to Assist Merging of Large-Scale Administrative Records. In: *American Political Science Review* 113.2, S. 353–371.

Fellegi, I. P. & A. B. Sunter (1969). A Theory for Record Linkage. In: *Journal of the American Statistical Association* 64.328, S. 1183–1210.

Gawronski, K. (2020). Konzeption eines Bildungsregisters in Deutschland. In: *WISTA - Wirtschaft und Statistik* 72.2, S. 37–45.

Giar, K., F. Hohlstein, M. Wipke & A. Scharnagl (2023). Konzeption eines Statistischen Bildungsverlaufsregisters in Deutschland – Entwicklungen bis 2023 und Ausgestaltungsoptionen. In: *WISTA - Wirtschaft und Statistik* 75.3, S. 51–62.

Gomatam, S., R. Carter, M. Ariet & G. Mitchell (2002). An Empirical Comparison of Record Linkage Procedures. In: *Statistics in Medicine* 21.10, S. 1485–1496.

Gupta, S., S. A. Roosz, J. A. LaBonte, V. Mucaj, J. O'Brien & A. Suresh (2023). "Linking of Tokenized Trial Data to Other Tokenized Data". US-Pat. 11,550,956 B1.

Hand, D. & P. Christen (2018). A Note on Using the F-Measure for Evaluating Record Linkage Algorithms. In: *Statistics and Computing* 28.3, S. 539–547.

Hand, D., P. Christen & N. Kirielle (2021). F*: An Interpretable Transformation of the F-measure. In: *Machine Learning* 110.3, S. 451–456.

Harron, K. L., J. C. Doidge, H. E. Knight, R. E. Gilbert, H. Goldstein, D. A. Cromwell & J. H. Van Der Meulen (2017). A Guide to Evaluating Linkage Quality for the Analysis of Linked Data. In: *International Journal of Epidemiology* 46.5, S. 1699–1710.

Hertweck, F., I. E. Isphording, S. H. Matthewes, K. Schneider & C. Katharina Spieß (2023). Bildungsdaten: Datenlücken durch ein Bildungsverlaufsregister schließen. In: *Wirtschaftsdienst* 103.11, S. 733–736.

Herzog, T. N., F. Scheuren & W. E. Winkler (2007). *Data Quality and Record Linkage Techniques*. New York: Springer.

Hosmer, D. W., S. Lemeshow & R. X. Sturdivant (2013). *Applied Logistic Regression*. 3. Aufl. Hoboken: Wiley.

INSPIRE Thematic Working Group Coordinate Reference Systems & Geographical Grid Systems (2014). *Data Specification on Geographical Grid Systems Technical Guidelines*. D2.8.I.2. Europäische Kommission.

Jaro, M. A. (1989). Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. In: *Journal of the American Statistical Association* 84.406, S. 414–420.

Kaalep, H.-J., F. Pirinen & S. Moshagen (2022). You Can't Suggest That?!: Comparisons and Improvements of Speller Error Models. In: *Nordlyd* 46.1, S. 125–139.

Kabisch, S. (2021). Wohnen in der Großwohnsiedlung. In: *Handbuch Wohnsoziologie*. Hrsg. von F. Eckardt & S. Meier. Wiesbaden: Springer, S. 295–312.

Kho, A. N. & S. Goel (2019). "Systems and Methods for Enabling Data De-Identification and Anonymous Data Linkage". US-Pat. 10,454,901 B2.

Kim, G. & R. L. Chambers (2012). Regression Analysis under Incomplete Linkage. In: *Computational Statistics & Data Analysis* 56.9, S. 2756–2770.

Koordinierungsstelle für IT-Standards (2022). *Datensatz für das Meldewesen: Einheitlicher Bundes-/Länderteil (DSMeld)*. Bremen: Koordinierungsstelle für IT-Standards.

Kristensen, T. G., J. Nielsen & C. N. S. Pedersen (2010). A Tree-Based Method for the Rapid Screening of Chemical Fingerprints. In: *Algorithms for Molecular Biology* 5.9.

Kukich, K. (1992). Techniques for Automatically Correcting Words in Text. In: *ACM Computing Surveys* 24.4, S. 377–439.

Kultusministerkonferenz (2004). *Kerndatensatz (KDS) für schulstatistische Individualdaten der Länder: Beschluss der Kultusministerkonferenz vom 08.05.2003*. KMK AL 115.

Kumar, U., V. Kumar & J. N. Kapur (1986). Normalized Measures of Entropy. In: *International Journal of General Systems* 12.1, S. 55–69.

Kvalsvig, A., S. Gibb & A. Teng (2019). *Linkage Error and Linkage Bias: A Guide for IDI Users*. Wellington: University of Otago.

Lenman, S. & H. Marmolin (1987). Naming Errors and Automatic Error Correction in Human-Computer Interaction. In: *Work with Display Units 86 : Selected Papers from the International Scientific Conference on Work With Display Units, Stockholm, Sweden, May 12-15, 1986*. Hrsg. von B. Knave & P.-G. Widebäck. Amsterdam: North-Holland, S. 838–846.

Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. In: *Soviet Physics – Doklady* 10.8, S. 707–710.

Linacre, R. (2023). *Probabilistic Record Linkage*. URL: https://www.robinlinacre.com/probabilistic_linkage/ (zuletzt abgerufen am 09. 02. 2024).

Lohr, S. L. (2021). *Sampling*. 3. Aufl. Boca Raton: CRC Press.

McCusker, M. E., R. D. Cress, M. Allen, A. Fernandez-Ami & R. Gandour-Edwards (2012). Feasibility of Linking Population-Based Cancer Registries and Cancer Center Biorepositories. In: *Biopreservation and Biobanking* 10.5, S. 416–420.

Meng, X.-L. & D. B. Rubin (1993). Maximum Likelihood Estimation via the ECM Algorithm: A General Framework. In: *Biometrika* 80.2, S. 267–278.

Mirel, L. B., D. M. Resnick, J. Aram & C. S. Cox (2022). A Methodological Assessment of Privacy Preserving Record Linkage Using Survey and Administrative Data. In: *Statistical Journal of the IAOS* 38.2, S. 413–421.

Moore, H. C., N. De Klerk, A. D. Keil, D. W. Smith, C. C. Blyth, P. Richmond & D. Lehmann (2012). Use of Data Linkage to Investigate the Aetiology of Acute Lower Respiratory Infection Hospitalisations in Children: Aetiology of Respiratory Infections. In: *Journal of Paediatrics and Child Health* 48.6, S. 520–528.

Mundelius, M. (2019). Der Kerndatensatz auf der Basis von Individualdatenerhebungen in der Schulstatistik. Von Summendaten zu Einzeldaten. In: *Bildungsforschung mit Daten der amtlichen Statistik*. Hrsg. von D. Fickermann & H. Weishaupt. Münster: Waxmann, S. 38–48.

NDR (2023). *Bundesweite Wohnungsnot: Auch in Hamburg spitzt sich die Lage zu*. URL: https://www.ndr.de/nachrichten/hamburg/Bundesweite-Wohnungsnot-Auch-in-Hamburg-spitzt-sich-die-Lage-zu,wohnungsnot252.html (zuletzt abgerufen am 13. 02. 2024).

OpenStreetMap Wiki (2023). *DE:Postleitzahlenänderungen in Deutschland*. URL: https://wiki.openstreetmap.org/wiki/DE:Postleitzahlen%C3%A4nderungen_in_Deutschland (zuletzt abgerufen am 10. 01. 2024).

Peterson, J. L. (1986). A Note on Undetected Typing Errors. In: *Communications of the ACM* 29.7, S. 633–637.

Pollock, J. J. & A. Zamora (1983). Collection and Characterization of Spelling Errors in Scientific and Scholarly Text. In: *Journal of the American Society for Information Science* 34.1, S. 51–58.

Pollock, J. J. & A. Zamora (1984). Automatic Spelling Correction in Scientific and Scholarly Text. In: *Communications of the ACM* 27.4, S. 358–368.

Randall, S., A. P. Brown, A. M. Ferrante & J. H. Boyd (2019). Privacy Preserving Linkage Using Multiple Dynamic Match Keys. In: *International Journal of Population Data Science* 4.1.

RatSWD (2023). *Aufbau eines Bildungsverlaufsregisters: Datenschutzkonform und forschungsfreundlich*. RatSWD Positionspapier.

Rothman, K. J. & S. Greenland (2014). Validity and Generalizability in Epidemiologic Studies. In: *Wiley StatsRef: Statistics Reference Online*. Hrsg. von R. S. Kenett, N. T. Longford, W. W. Piegorsch & F. Ruggeri. Wiley.

Schnell, R. (2016). Privacy-preserving Record Linkage. In: *Methodological Developments in Data Linkage*. Hrsg. von K. Harron, H. Goldstein & C. Dibben. Chichester: Wiley, S. 201–225.

Schnell, R. (2019a). *Eignung von Personenmerkmalen als Datengrundlage zur Verknüpfung von Registerinformationen im Integrierten Registerzensus*. WP-GRLC-2019-01. Duisburg: DuEPublico.

Schnell, R. (2019b). *Survey-Interviews: Methoden standardisierter Befragungen*. 2. Aufl. Wiesbaden: Springer.

Schnell, R. (2022). *Verknüpfung von Bildungsdaten in einem Bildungsregister mittels Record-Linkage auf Basis von Personenmerkmalen*. WP-GRLC-2022-03. Duisburg: DuEPublico.

Schnell, R., T. Bachteler & J. Reiher (2009). Privacy-Preserving Record Linkage Using Bloom Filters. In: *BMC Medical Informatics and Decision Making* 9.41.

Schnell, R., T. Bachteler & J. Reiher (2011). *A Novel Error-Tolerant Anonymous Linking Code*. WP-GRLC-2011-02.

Schnell, R., J. Klingwort & J. M. Farrow (2021). Locational Privacy-Preserving Distance Computations with Intersecting Sets of Randomly Labeled Grid Points. In: *International Journal of Health Geographics* 20.1, S. 14.

Schnell, R. & S. V. Weiand (2023). Microsimulation of an Educational Attainment Register to Predict Future Record Linkage Quality. In: *International Journal of Population Data Science* 8.1.

Shannon, C. E. (1948). A Mathematical Theory of Communication. In: *The Bell System Technical Journal* 27, S. 379–423, 623–656.

Shipsey, R. & J. Plachta (2021). *Linking with Anonymised Data – How Not to Make a Hash of It*. Office for National Statistics. URL: https://www.gov.uk/government/publications/joined-up-data-in-government-the-future-of-data-linking-methods/linking-with-anonymised-data-how-not-to-make-a-hash-of-it (zuletzt abgerufen am 31. 01. 2024).

Statistikamt Nord (2023). *Bevölkerung mit Migrationshintergrund in den Hamburger Stadtteilen 2022: Auswertung auf Basis des Melderegisters*. A I 10 - j 22 HH. Hamburg: Statistisches Amt für Hamburg und Schleswig-Holstein.

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

Statistisches Bundesamt (2023). *Statistik der allgemeinbildenden Schulen: 08/2022-07/2023*. Qualitätsbericht. Wiesbaden: Statistisches Bundesamt.

Tagliacozzo, R., M. Kochen & L. Rosenberg (1970). Orthographic Error Patterns of Author Names in Catalog Searches. In: *Journal of Library Automation* 3.2, S. 93–101.

Tromp, M., A. C. Ravelli, G. J. Bonsel, A. Hasman & J. B. Reitsma (2011). Results from Simulated Data Sets: Probabilistic Record Linkage Outperforms Deterministic Record Linkage. In: *Journal of Clinical Epidemiology* 64.5, S. 565–572.

Wang, J., Y. Liu, P. Li, Z. Lin, S. Sindakis & S. Aggarwal (2023). Overview of Data Quality: Examining the Dimensions, Antecedents, and Impacts of Data Quality. In: *Journal of the Knowledge Economy*.

Weiand, S. V. (2022). *Vergleich von Record-Linkage Methoden anhand der Mikrosimulation eines bundesweiten Schülerregisters*. WP-GRLC-2022-04. Duisburg: DuEPublico.

Winkler, W. E. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. In: *Proceedings of the Section on Survey Research Methods*. Bd. 13. American Statistical Association / U.S. Bureau of the Census, S. 354–357.

Wurdeman, K. (1993). Quality of Data Keying for Major Operations of the 1990 Census. In: *Proceedings of the Section on Survey Research Methods, American Statistical Association* 16, S. 312–317.

Yancey, W. E. (2005). *Evaluating String Comparator Performance for Record Linkage*. Statistics #2005-05. Washington: U.S. Census Bureau.

Anhang A.

Weiterführende Ergebnisse

Tabelle A.1 stellt die Ergebnisse aller in Abschnitt 5.3 besprochenen Record-Linkage-Verfahren für alle 40 Szenarien dar. In den Tabellen A.2, A.3, A.4 und A.5 werden die Logits der logistischen Regression aus Abschnitt 5.4.4 für die Varianten des probabilistischen Record-Linkage, die CLKs und das Datavant-Verfahren dargestellt.

Tabelle A.1.: Linkage-Ergebnisse und Bias aller getesteten Szenarien und Verfahren. Für Precision, Recall und F^* ist jeweils das Maximum innerhalb eines Szenarios dick gedruckt (Sz. = Szenario; Schw. = Schwellenwert; n_k = Anzahl verwendeter Matchkeys; Prec. = Precision; Sf. = Schulform; Quelle = Daten stammen von einer Schule; Mig. = Migrant; Mh. = Migrationshintergrund; Deu. = Deutsch; Ges. = Geschlecht; um. = Anteil richtig verknüpfter umgezogener Personen; d = Cohens's d; h = Cohen's h; V = Cramer's V; MMK = Multiple Matchkeys; DTV = Datavant; ECM(D) = ECM mit Dice-Modifikation).

Sz.	QIDs	Methode	Schw.	n_k	TP	FP	FN	Prec.	Recall	F^*	d_{Alter}	$V_{Sf.}$	h_{Quelle}	$h_{Mig.}$	$h_{Mh.}$	$h_{Deu.}$	$h_{Ges.}$	um.(%)
1	Vorname	Exact			159707	0	22678	**1.0000**	0.8757	0.8757	0.04	0.03	-0.05	-0.04	-0.02	0.05	0.00	0.00
	Nachname	MMK	0.70	3	162801	904	19584	0.9945	0.8926	0.8882	0.02	0.02	-0.02	-0.04	0.00	0.03	0.00	0.36
	Geschlecht		0.90	5	169967	11786	12418	0.9352	0.9319	0.8753	-0.08	0.03	0.08	-0.02	-0.01	0.02	0.00	26.36
	Staatsbürgers.		0.98	10	168335	14755	14050	0.9194	0.9230	0.8539	-0.1	0.04	0.10	-0.01	-0.01	0.01	0.00	15.92
	Tag, Monat	CLK	0.80		166411	79	15974	0.9995	0.9124	0.9120	0.01	0.01	-0.02	-0.03	0.00	0.02	0.00	8.41
	Jahr		0.85		165714	17	16671	0.9999	0.9086	0.9085	0.02	0.02	-0.03	-0.03	0.00	0.02	0.00	7.71
	Postleitzahl	DTV		8	181785	96	600	0.9995	**0.9967**	**0.9962**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	98.69
	Straße	ECM	0.80		181395	184	990	0.9990	0.9946	0.9936	0.00	0.00	0.00	0.00	0.00	0.00	0.00	94.48
	Hausnummer		0.90		181391	184	994	0.9990	0.9945	0.9935	0.00	0.00	0.00	0.00	0.00	0.00	0.00	94.45
		ECM(D)	0.80		181520	186	865	0.9990	0.9953	0.9942	0.00	0.00	0.00	0.00	0.00	0.00	0.00	95.15
			0.90		181520	186	865	0.9990	0.9953	0.9942	0.00	0.00	0.00	0.00	0.00	0.00	0.00	95.15
2	Vorname	Exact			164396	0	17989	**1.0000**	0.9014	0.9014	0.04	0.03	-0.04	-0.04	-0.03	0.05	0.00	25.90
	Nachname	MMK	0.70	3	178334	5351	4051	0.9709	0.9778	0.9499	-0.04	0.02	0.04	0.00	-0.01	0.01	0.00	85.75
	Geschlecht		0.90	2	179634	4263	2751	0.9768	0.9849	0.9624	-0.03	0.01	0.03	-0.01	0.00	0.01	0.00	96.36
	Staatsbürgers.		0.98	6	180851	9902	1534	0.9481	0.9916	0.9405	-0.08	0.04	0.09	0.00	-0.01	0.01	0.00	98.50
	Tag, Monat	CLK	0.80		179818	33	2567	0.9998	0.9859	0.9857	0.01	0.00	-0.01	-0.01	0.00	0.01	0.00	94.58
	Jahr		0.85		177359	9	5026	0.9999	0.9724	0.9724	0.02	0.01	-0.02	-0.01	-0.01	0.01	0.00	85.32
	Postleitzahl	DTV		11	180968	33	1417	0.9998	0.9922	0.9921	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	98.69
		ECM	0.80		181644	66	741	0.9996	0.9959	0.9956	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	98.08
			0.90		181557	40	828	0.9998	0.9955	0.9952	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	98.07
		ECM(D)	0.80		181708	50	677	0.9997	**0.9963**	**0.9960**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	98.67
			0.90		181215	45	1170	0.9998	0.9936	0.9933	0.00	0.00	0.00	0.00	0.00	0.00	0.00	98.56
3	Vorname	Exact			160623	0	21762	**1.0000**	0.8807	0.8807	0.04	0.03	-0.04	-0.04	-0.02	0.05	0.00	3.15
	Nachname	MMK	0.70	5	182018	5642	367	0.9699	**0.9980**	0.9680	-0.05	0.02	0.05	0.00	0.00	0.00	0.00	98.91
	Geschlecht		0.90	4	180692	6630	1693	0.9646	0.9907	0.9560	-0.06	0.02	0.06	0.00	-0.01	0.01	0.00	93.83
	Staatsbürgers.		0.98	4	180729	6596	1656	0.9648	0.9909	0.9563	-0.05	0.02	0.06	0.00	-0.01	0.01	0.00	93.95
	Tag, Monat	CLK	0.80		180804	223	1581	0.9988	0.9913	0.9901	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	97.86
	Jahr		0.85		179816	46	2569	0.9997	0.9859	0.9857	0.01	0.00	-0.01	-0.01	0.00	0.01	0.00	96.21
	100m-Gitter	DTV		11	181780	68	605	0.9996	0.9967	0.9963	0.00	0.00	0.00	0.00	0.00	0.00	0.00	98.73
		ECM	0.80		181675	58	710	0.9997	0.9961	0.9958	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	97.86
			0.90		181645	58	740	0.9997	0.9959	0.9956	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	97.77
		ECM(D)	0.80		181992	81	393	0.9996	0.9978	**0.9974**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	98.76
			0.90		181954	81	431	0.9996	0.9976	0.9972	0.00	0.00	0.00	0.00	0.00	0.00	0.00	98.53
4	Vorname	Exact			162427	0	19958	**1.0000**	0.8906	0.8906	0.04	0.03	-0.04	-0.04	-0.03	0.05	0.00	14.03
	Nachname	MMK	0.70	3	179462	4106	2923	0.9776	0.9840	0.9623	-0.03	0.01	0.04	-0.01	0.00	0.01	0.00	88.10
	Geschlecht		0.90	3	181229	6074	1156	0.9676	0.9937	0.9616	-0.05	0.02	0.05	0.00	-0.01	0.01	0.00	96.35
	Staatsbürgers.		0.98	6	180850	9669	1535	0.9492	0.9916	0.9417	-0.08	0.04	0.09	0.00	-0.01	0.01	0.00	98.50
	Tag, Monat	CLK	0.80		180768	178	1617	0.9990	0.9911	0.9902	0.01	0.00	0.00	-0.01	0.00	0.00	0.00	97.80
	Jahr		0.85		179714	41	2671	0.9998	0.9854	0.9851	0.01	0.00	-0.01	-0.01	0.00	0.01	0.00	95.79
	1km-Gitter	DTV		7	181133	64	1252	0.9996	0.9931	0.9928	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	98.72
		ECM	0.80		181608	59	777	0.9997	0.9957	0.9954	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	97.87
			0.90		181608	59	777	0.9997	0.9957	0.9954	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	97.87
		ECM(D)	0.80		181700	46	685	0.9997	**0.9962**	**0.9960**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	98.58
			0.90		181700	46	685	0.9997	**0.9962**	**0.9960**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	98.58

Tabelle A.1.: Fortsetzung der Linkage-Ergebnisse aller getesteten Szenarien und Verfahren.

Sz.	QIDs	Methode	Schw.	n_k	TP	FP	FN	Prec.	Recall	F^*	d_{Alter}	$V_{Sf.}$	h_{Quelle}	$h_{Mig.}$	$h_{Mh.}$	$h_{Deu.}$	$h_{Ges.}$	um.(%)
5	Vorname	Exact			175725	0	6660	**1.0000**	0.9635	0.9635	0.03	0.01	-0.02	-0.01	-0.03	0.03	0.00	93.72
	Nachname	MMK	0.70	5	178431	6985	3954	0.9623	0.9783	0.9422	-0.05	0.02	0.06	0.01	-0.03	0.01	0.00	96.62
	Geschlecht		0.90	2	179799	4275	2586	0.9768	0.9858	0.9632	-0.03	0.01	0.03	-0.01	0.00	0.01	0.00	97.32
	Staatsbürgers.		0.98	2	179799	4275	2586	0.9768	0.9858	0.9632	-0.03	0.01	0.03	-0.01	0.00	0.01	0.00	97.32
	Tag, Monat	CLK	0.80		179939	92	2446	0.9995	0.9866	0.9861	0.01	0.00	-0.01	-0.01	0.00	0.01	0.00	97.61
	Jahr		0.85		178952	36	3433	0.9998	0.9812	0.9810	0.01	0.00	-0.01	-0.01	-0.01	0.01	0.00	96.73
		DTV		10	180968	33	1417	0.9998	0.9922	0.9921	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	98.69
		ECM	0.80		181808	9173	577	0.9520	**0.9968**	0.9491	-0.08	0.04	0.10	-0.01	-0.02	0.02	0.00	99.57
			0.90		180662	1988	1723	0.9891	0.9906	0.9799	-0.01	0.01	0.02	0.00	0.00	0.01	0.00	98.46
		ECM(D)	0.80		181136	64	1249	0.9996	0.9932	**0.9928**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	98.81
			0.90		181064	44	1321	0.9998	0.9928	0.9925	0.00	0.00	-0.01	0.00	0.00	0.00	0.00	98.72
6	Vorname	Exact			162360	0	20025	**1.0000**	0.8902	0.8902	0.03	0.03	-0.05	-0.04	0.00	0.03	0.00	0.00
	Nachname	MMK	0.70	4	167658	10065	14727	0.9434	0.9193	0.8712	-0.07	0.03	0.07	-0.03	-0.01	0.03	0.00	16.17
	Geschlecht		0.90	6	167681	9149	14704	0.9483	0.9194	0.8755	-0.07	0.03	0.06	-0.03	-0.01	0.02	0.00	16.48
	Tag, Monat		0.98	7	169607	3753	12778	0.9784	0.9299	0.9112	-0.02	0.01	0.02	-0.03	0.00	0.02	0.00	26.09
	Jahr	CLK	0.80		166364	65	16021	0.9996	0.9122	0.9118	0.01	0.01	-0.02	-0.03	0.00	0.02	0.00	8.29
	Postleitzahl		0.85		165677	17	16708	0.9999	0.9084	0.9083	0.02	0.02	-0.03	-0.03	0.00	0.02	0.00	7.71
	Straße	DTV		7	181683	60	702	0.9997	0.9962	0.9958	0.00	0.00	0.00	0.00	0.00	0.00	0.00	98.58
	Hausnummer	ECM	0.80		181934	85	451	0.9995	**0.9975**	**0.9971**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	97.62
			0.90		181930	85	455	0.9995	0.9975	0.9970	0.00	0.00	0.00	0.00	0.00	0.00	0.00	97.60
		ECM(D)	0.80		181740	338	645	0.9981	0.9965	0.9946	0.00	0.00	0.00	0.00	0.00	0.00	0.00	96.48
			0.90		170360	336	12025	0.9980	0.9341	0.9324	0.01	0.01	-0.02	-0.03	0.00	0.02	0.00	28.09
7	Vorname	Exact			167188	0	15197	**1.0000**	0.9167	0.9167	0.03	0.02	-0.04	-0.04	0.00	0.02	0.00	26.49
	Nachname	MMK	0.70	2	179670	4265	2715	0.9768	0.9851	0.9626	-0.03	0.01	0.03	-0.01	0.00	0.01	0.00	96.36
	Geschlecht		0.90	2	179670	4265	2715	0.9768	0.9851	0.9626	-0.03	0.01	0.03	-0.01	0.00	0.01	0.00	96.36
	Tag, Monat		0.98	7	181229	2807	1156	0.9847	0.9937	0.9786	-0.02	0.01	0.02	0.00	0.00	0.00	0.00	95.77
	Jahr	CLK	0.80		179625	26	2760	0.9999	0.9849	0.9847	0.01	0.00	-0.01	-0.01	0.00	0.00	0.00	93.60
	Postleitzahl		0.85		177433	10	4952	0.9999	0.9728	0.9728	0.02	0.01	-0.02	-0.01	0.00	0.01	0.00	82.71
		DTV		10	180895	22	1490	0.9999	0.9918	0.9917	0.00	0.00	-0.01	-0.01	0.00	0.00	0.00	98.57
		ECM	0.80		181672	46	713	0.9997	0.9961	0.9958	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	98.09
			0.90		181672	46	713	0.9997	0.9961	0.9958	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	98.09
		ECM(D)	0.80		181746	54	639	0.9997	**0.9965**	**0.9962**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	98.68
			0.90		181165	45	1220	0.9998	0.9933	0.9931	0.00	0.00	0.00	0.00	0.00	0.00	0.00	98.53
8	Vorname	Exact			163340	0	19045	**1.0000**	0.8956	0.8956	0.03	0.02	-0.04	-0.04	0.00	0.02	0.00	3.29
	Nachname	MMK	0.70	4	180706	6632	1679	0.9646	0.9908	0.9560	-0.06	0.02	0.06	0.00	-0.01	0.01	0.00	93.89
	Geschlecht		0.90	4	180706	6632	1679	0.9646	0.9908	0.9560	-0.06	0.02	0.06	0.00	-0.01	0.01	0.00	93.89
	Tag, Monat		0.98	3	181214	5401	1171	0.9711	0.9936	0.9650	-0.05	0.02	0.05	-0.01	-0.01	0.01	0.00	96.31
	Jahr	CLK	0.80		180758	166	1627	0.9991	0.9911	0.9902	0.01	0.00	0.00	0.00	0.00	0.00	0.00	97.98
	100m-Gitter		0.85		179963	32	2422	0.9998	0.9867	0.9865	0.01	0.00	-0.01	-0.01	0.00	0.00	0.00	96.64
		DTV		10	181759	57	626	0.9997	0.9966	0.9963	0.00	0.00	0.00	0.00	0.00	0.00	0.00	98.62
		ECM	0.80		181646	59	739	0.9997	0.9959	0.9956	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	97.77
			0.90		181646	59	739	0.9997	0.9959	0.9956	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	97.77
		ECM(D)	0.80		182005	92	380	0.9995	**0.9979**	**0.9974**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	98.53
			0.90		182004	92	381	0.9995	0.9979	0.9974	0.00	0.00	0.00	0.00	0.00	0.00	0.00	98.53

Tabelle A.1.: Fortsetzung der Linkage-Ergebnisse aller getesteten Szenarien und Verfahren.

Sz.	QIDs	Methode	Schw.	n_k	TP	FP	FN	Prec.	Recall	F^*	d_{Alter}	$V_{Sf.}$	h_{Quelle}	$h_{Mig.}$	$h_{Mh.}$	$h_{Deu.}$	$h_{Ges.}$	um.(%)
9	Vorname	Exact			165192	0	17193	**1.0000**	0.9057	0.9057	0.03	0.02	-0.04	-0.04	0.00	0.02	0.00	14.45
	Nachname	MMK	0.70	2	179648	4263	2737	0.9768	0.9850	0.9625	-0.03	0.01	0.03	-0.01	0.00	0.01	0.00	96.22
	Geschlecht		0.90	2	179648	4263	2737	0.9768	0.9850	0.9625	-0.03	0.01	0.03	-0.01	0.00	0.01	0.00	96.22
	Tag, Monat		0.98	7	180579	9499	1806	0.9500	0.9901	0.9411	-0.08	0.04	0.08	0.00	-0.01	0.01	0.00	97.72
	Jahr	CLK	0.80		180729	142	1656	0.9992	0.9909	0.9901	0.01	0.00	0.00	-0.01	0.00	0.00	0.00	97.94
	1km-Gitter		0.85		179848	33	2537	0.9998	0.9861	0.9859	0.01	0.00	-0.01	-0.01	0.00	0.00	0.00	96.17
		DTV		6	181113	53	1272	0.9997	0.9930	0.9927	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	98.61
		ECM	0.80		181651	60	734	0.9997	0.9960	0.9956	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	97.88
			0.90		181637	36	748	0.9998	0.9959	0.9957	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	97.88
		ECM(D)	0.80		181732	47	653	0.9997	**0.9964**	**0.9962**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	98.58
			0.90		181730	46	655	0.9997	0.9964	0.9962	0.00	0.00	0.00	0.00	0.00	0.00	0.00	98.58
10	Vorname	Exact			178838	0	3547	**1.0000**	0.9806	0.9806	0.02	0.01	-0.02	-0.01	0.00	0.01	0.00	96.23
	Nachname	MMK	0.70	2	179839	4276	2546	0.9768	0.9860	0.9635	-0.03	0.01	0.03	-0.01	0.00	0.01	0.00	97.36
	Geschlecht		0.90	2	179839	4276	2546	0.9768	0.9860	0.9635	-0.03	0.01	0.03	-0.01	0.00	0.01	0.00	97.36
	Tag, Monat		0.98	2	179839	4276	2546	0.9768	0.9860	0.9635	-0.03	0.01	0.03	-0.01	0.00	0.01	0.00	97.36
	Jahr	CLK	0.80		180051	81	2334	0.9996	0.9872	0.9868	0.01	0.00	-0.01	-0.01	0.00	0.00	0.00	97.67
			0.85		179719	30	2666	0.9998	0.9854	0.9852	0.01	0.00	-0.01	-0.01	0.00	0.00	0.00	97.31
		DTV		4	180893	22	1492	0.9999	0.9918	0.9917	0.00	0.00	-0.01	-0.01	0.00	0.00	0.00	98.57
		ECM	0.80		180525	22	1860	0.9999	0.9898	0.9897	0.01	0.00	-0.01	-0.01	0.00	0.00	0.00	98.23
			0.90		180449	1	1936	1.0000	0.9894	0.9894	0.01	0.00	-0.01	-0.01	0.00	0.00	0.00	98.14
		ECM(D)	0.80		181074	45	1311	0.9998	**0.9928**	**0.9926**	0.00	0.00	-0.01	0.00	0.00	0.00	0.00	98.72
			0.90		181062	41	1323	0.9998	0.9927	0.9925	0.00	0.00	-0.01	0.00	0.00	0.00	0.00	98.72
11	Vorname	Exact			159757	0	22628	**1.0000**	0.8759	0.8759	0.04	0.03	-0.05	-0.04	-0.02	0.05	0.00	0.00
	Nachname	MMK	0.70	3	162965	912	19420	0.9944	0.8935	0.8891	0.02	0.01	-0.02	-0.04	0.00	0.03	0.00	0.36
	Geschlecht		0.90	2	159877	902	22508	0.9944	0.8766	0.8723	0.03	0.02	-0.03	-0.04	-0.03	0.05	0.00	0.04
	Staatsbürgers.		0.98	7	167188	8723	15197	0.9504	0.9167	0.8748	-0.06	0.03	0.06	-0.03	-0.01	0.02	0.00	13.56
	Monat	CLK	0.80		166308	152	16077	0.9991	0.9119	0.9111	0.01	0.01	-0.02	-0.03	0.00	0.02	0.00	8.24
	Jahr		0.85		165614	32	16771	0.9998	0.9080	0.9079	0.02	0.02	-0.03	-0.03	0.00	0.02	0.00	7.66
	Postleitzahl	DTV		1	180035	51	2350	0.9997	**0.9871**	**0.9868**	0.01	0.00	-0.01	-0.01	0.00	0.01	0.00	93.81
	Straße	ECM	0.80		170267	160	12118	0.9991	0.9336	0.9327	0.01	0.01	-0.02	-0.03	0.00	0.02	0.00	27.60
	Hausnummer		0.90		170176	157	12209	0.9991	0.9331	0.9323	0.01	0.01	-0.02	-0.03	0.00	0.02	0.00	27.08
		ECM(D)	0.80		169947	3149	12438	0.9818	0.9318	0.9160	0.00	0.00	0.00	-0.02	0.00	0.01	0.00	25.71
			0.90		169914	2480	12471	0.9856	0.9316	0.9191	0.00	0.00	-0.01	-0.02	0.00	0.02	0.00	25.47
12	Vorname	Exact			164453	0	17932	**1.0000**	0.9017	0.9017	0.04	0.03	-0.04	-0.04	-0.03	0.05	0.00	25.93
	Nachname	MMK	0.70	4	180387	6602	1998	0.9647	0.9890	0.9545	-0.05	0.03	0.06	0.01	-0.02	0.01	0.00	97.40
	Geschlecht		0.90	2	179637	4263	2748	0.9768	0.9849	0.9624	-0.03	0.01	0.03	-0.01	0.00	0.01	0.00	96.36
	Staatsbürgers.		0.98	2	179637	4263	2748	0.9768	0.9849	0.9624	-0.03	0.01	0.03	-0.01	0.00	0.01	0.00	96.36
	Monat	CLK	0.80		179333	82	3052	0.9995	0.9833	0.9828	0.01	0.00	-0.01	-0.01	0.00	0.01	0.00	92.11
	Jahr		0.85		176512	25	5873	0.9999	0.9678	0.9677	0.02	0.01	-0.02	-0.02	0.00	0.01	0.00	80.11
	Postleitzahl	DTV		2	176012	14	6373	0.9999	0.9651	0.9650	0.03	0.01	-0.02	-0.01	-0.03	0.03	0.00	94.03
		ECM	0.80		181648	7347	737	0.9611	**0.9960**	0.9574	-0.07	0.03	0.07	-0.01	-0.01	0.02	0.00	98.38
			0.90		181070	2001	1315	0.9891	0.9928	0.9820	-0.02	0.01	0.02	0.00	0.00	0.01	0.00	97.97
		ECM(D)	0.80		180844	74	1541	0.9996	0.9916	**0.9911**	0.01	0.00	-0.01	0.00	0.00	0.01	0.00	96.54
			0.90		180835	72	1550	0.9996	0.9915	0.9911	0.01	0.00	-0.01	0.00	0.00	0.01	0.00	96.54

Tabelle A.1.: Fortsetzung der Linkage-Ergebnisse aller getesteten Szenarien und Verfahren.

Sz.	QIDs	Methode	Schw.	n_k	TP	FP	FN	Prec.	Recall	F^*	d_{Alter}	$V_{Sf.}$	h_{Quelle}	$h_{Mig.}$	$h_{Mh.}$	$h_{Deu.}$	$h_{Ges.}$	um.(%)
13	Vorname	Exact			160675	0	21710	**1.0000**	0.8810	0.8810	0.04	0.03	-0.04	-0.04	-0.02	0.05	0.00	3.15
	Nachname	MMK	0.70	5	181933	4551	452	0.9756	**0.9975**	0.9732	-0.04	0.02	0.04	0.01	0.00	-0.01	0.00	98.55
	Geschlecht		0.90	4	180692	6630	1693	0.9646	0.9907	0.9560	-0.06	0.02	0.06	0.00	-0.01	0.01	0.00	93.83
	Staatsbürgers.		0.98	4	180692	6630	1693	0.9646	0.9907	0.9560	-0.06	0.02	0.06	0.00	-0.01	0.01	0.00	93.83
	Monat	CLK	0.80		180644	462	1741	0.9974	0.9905	0.9880	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	97.44
	Jahr		0.85		179522	116	2863	0.9994	0.9843	0.9837	0.01	0.00	-0.01	-0.01	0.00	0.01	0.00	94.34
	100m-Gitter	DTV		2	180060	14	2325	0.9999	0.9873	0.9872	0.01	0.00	-0.01	-0.01	0.00	0.01	0.00	94.26
		ECM	0.80		181811	3757	574	0.9798	0.9969	0.9767	-0.03	0.02	0.04	0.00	-0.01	0.01	0.00	98.28
			0.90		181732	2165	653	0.9882	0.9964	0.9847	-0.02	0.01	0.02	0.00	0.00	0.00	0.00	97.80
		ECM(D)	0.80		181315	72	1070	0.9996	0.9941	**0.9937**	0.00	0.00	0.00	0.00	0.00	0.01	0.00	96.03
			0.90		181315	72	1070	0.9996	0.9941	**0.9937**	0.00	0.00	0.00	0.00	0.00	0.01	0.00	96.03
14	Vorname	Exact			162480	0	19905	**1.0000**	0.8909	0.8909	0.04	0.03	-0.04	-0.04	-0.03	0.05	0.00	14.04
	Nachname	MMK	0.70	4	181621	5306	764	0.9716	0.9958	0.9677	-0.05	0.02	0.05	0.00	-0.01	0.01	0.00	97.87
	Geschlecht		0.90	3	179567	4920	2818	0.9733	0.9845	0.9587	-0.04	0.02	0.04	0.00	0.00	0.00	0.00	90.61
	Staatsbürgers.		0.98	2	179613	4261	2772	0.9768	0.9848	0.9623	-0.03	0.01	0.03	-0.01	0.00	0.01	0.00	96.21
	Monat	CLK	0.80		180629	447	1756	0.9975	0.9904	0.9880	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	97.44
	Jahr		0.85		179444	102	2941	0.9994	0.9839	0.9833	0.01	0.00	-0.01	-0.01	0.00	0.01	0.00	94.15
	1km-Gitter	DTV		1	178936	12	3449	0.9999	0.9811	0.9810	0.02	0.01	-0.02	-0.01	0.00	0.01	0.00	94.48
		ECM	0.80		181660	6570	725	0.9651	**0.9960**	0.9614	-0.06	0.03	0.07	0.00	-0.01	0.01	0.00	98.35
			0.90		181587	3982	798	0.9785	0.9956	0.9744	-0.04	0.02	0.04	-0.01	-0.01	0.01	0.00	97.91
		ECM(D)	0.80		180876	74	1509	0.9996	0.9917	**0.9913**	0.00	0.00	0.00	0.00	0.00	0.01	0.00	96.33
			0.90		180809	71	1576	0.9996	0.9914	0.9910	0.01	0.00	-0.01	0.00	0.00	0.01	0.00	96.32
15	Vorname	Exact			175774	3	6611	**1.0000**	0.9638	0.9637	0.03	0.01	-0.02	-0.01	-0.03	0.03	0.00	93.82
	Nachname	MMK	0.70	3	178295	4754	4090	0.9740	0.9776	0.9527	-0.03	0.01	0.03	0.00	-0.02	0.01	0.00	96.21
	Geschlecht		0.90	2	179781	4275	2604	0.9768	0.9857	0.9631	-0.03	0.01	0.03	-0.01	0.00	0.01	0.00	97.32
	Staatsbürgers.		0.98	2	179781	4275	2604	0.9768	0.9857	0.9631	-0.03	0.01	0.03	-0.01	0.00	0.01	0.00	97.32
	Monat	CLK	0.80		179559	288	2826	0.9984	0.9845	0.9830	0.01	0.00	-0.01	-0.01	0.00	0.01	0.00	97.30
	Jahr		0.85		178518	127	3867	0.9993	0.9788	0.9781	0.02	0.00	-0.01	-0.01	-0.01	0.01	0.00	96.34
		DTV		2	175990	14	6395	0.9999	0.9649	0.9649	0.03	0.01	-0.02	-0.01	-0.03	0.03	0.00	94.01
		ECM	0.80		181150	3608	1235	0.9805	**0.9932**	0.9740	-0.03	0.02	0.04	-0.01	-0.01	0.01	0.00	98.85
			0.90		180497	1964	1888	0.9892	0.9896	0.9791	-0.01	0.00	0.02	0.00	0.00	0.01	0.00	98.20
		ECM(D)	0.80		181124	110	1261	0.9994	0.9931	**0.9925**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	98.81
			0.90		181124	110	1261	0.9994	0.9931	**0.9925**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	98.81
16	Vorname	Exact			162413	0	19972	**1.0000**	0.8905	0.8905	0.03	0.02	-0.05	-0.04	0.00	0.03	0.00	0.00
	Nachname	MMK	0.70	2	162518	913	19867	0.9944	0.8911	0.8866	0.02	0.02	-0.03	-0.04	0.00	0.03	0.00	0.04
	Geschlecht		0.90	4	167658	10065	14727	0.9434	0.9193	0.8712	-0.07	0.03	0.07	-0.03	-0.01	0.03	0.00	16.17
	Monat		0.98	6	167681	9150	14704	0.9483	0.9194	0.8755	-0.07	0.03	0.06	-0.03	-0.01	0.02	0.00	16.48
	Jahr	CLK	0.80		166245	131	16140	0.9992	0.9115	0.9109	0.01	0.01	-0.02	-0.03	0.00	0.02	0.00	8.13
	Postleitzahl		0.85		165577	30	16808	0.9998	0.9078	0.9077	0.02	0.02	-0.03	-0.03	0.00	0.02	0.00	7.66
	Straße	ECM	0.80		170278	290	12107	0.9983	**0.9336**	**0.9321**	0.01	0.01	-0.02	-0.03	0.00	0.02	0.00	27.43
	Hausnummer		0.90		170274	289	12111	0.9983	0.9336	0.9321	0.01	0.01	-0.02	-0.03	0.00	0.02	0.00	27.42
		ECM(D)	0.80		169869	3168	12516	0.9817	0.9314	0.9155	0.00	0.00	-0.01	-0.02	0.00	0.01	0.00	25.22
			0.90		169869	3168	12516	0.9817	0.9314	0.9155	0.00	0.00	-0.01	-0.02	0.00	0.01	0.00	25.22
17	Vorname	Exact			167248	0	15137	**1.0000**	0.9170	0.9170	0.03	0.02	-0.04	-0.04	0.00	0.02	0.00	26.52
	Nachname	MMK	0.70	2	179673	4265	2712	0.9768	0.9851	0.9626	-0.03	0.01	0.03	-0.01	0.00	0.01	0.00	96.36
	Geschlecht		0.90	2	179673	4265	2712	0.9768	0.9851	0.9626	-0.03	0.01	0.03	-0.01	0.00	0.01	0.00	96.36
	Monat		0.98	5	180566	7581	1819	0.9597	0.9900	0.9505	-0.06	0.03	0.06	0.00	-0.01	0.01	0.00	97.99
	Jahr	CLK	0.80		178928	73	3457	0.9996	0.9810	0.9807	0.01	0.01	-0.01	-0.01	0.00	0.01	0.00	89.83
	Postleitzahl		0.85		176422	22	5963	0.9999	0.9673	0.9672	0.02	0.01	-0.02	-0.02	0.00	0.01	0.00	76.77
		ECM	0.80		181102	91	1283	0.9995	0.9930	0.9925	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	97.84
			0.90		180530	29	1855	0.9998	0.9898	0.9897	0.01	0.00	-0.01	-0.01	0.00	0.00	0.00	97.68
		ECM(D)	0.80		181213	91	1172	0.9995	**0.9936**	**0.9931**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	98.61
			0.90		181101	87	1284	0.9995	0.9930	0.9925	0.00	0.00	0.00	0.00	0.00	0.00	0.00	98.58

Tabelle A.1.: Fortsetzung der Linkage-Ergebnisse aller getesteten Szenarien und Verfahren.

Sz.	QIDs	Methode	Schw.	n_k	TP	FP	FN	Prec.	Recall	F^*	d_{Alter}	$V_{Sf.}$	h_{Quelle}	$h_{Mig.}$	$h_{Mh.}$	$h_{Deu.}$	$h_{Ges.}$	um.(%)
18	Vorname	Exact			163395	0	18990	**1.0000**	0.8959	0.8959	0.03	0.02	-0.04	-0.04	0.00	0.02	0.00	3.29
	Nachname	MMK	0.70	4	180706	6632	1679	0.9646	0.9908	0.9560	-0.06	0.02	0.06	0.00	-0.01	0.01	0.00	93.89
	Geschlecht		0.90	4	180706	6632	1679	0.9646	0.9908	0.9560	-0.06	0.02	0.06	0.00	-0.01	0.01	0.00	93.89
	Monat		0.98	3	180650	6406	1735	0.9658	0.9905	0.9569	-0.05	0.02	0.06	-0.01	-0.01	0.01	0.00	94.04
	Jahr	CLK	0.80		180612	398	1773	0.9978	0.9903	0.9881	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	97.58
	100m-Gitter		0.85		179516	111	2869	0.9994	0.9843	0.9837	0.01	0.00	-0.01	-0.01	0.00	0.00	0.00	94.24
		DTV		1	164727	2	17658	1.0000	0.9032	0.9032	0.02	0.02	-0.03	-0.04	0.00	0.02	0.00	3.39
		ECM	0.80		181655	87	730	0.9995	0.9960	0.9955	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	97.61
			0.90		181655	87	730	0.9995	0.9960	0.9955	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	97.61
		ECM(D)	0.80		181777	90	608	0.9995	**0.9967**	**0.9962**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	98.58
			0.90		181777	90	608	0.9995	**0.9967**	**0.9962**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	98.58
19	Vorname	Exact			165248	0	17137	**1.0000**	0.9060	0.9060	0.03	0.02	-0.04	-0.04	0.00	0.02	0.00	14.46
	Nachname	MMK	0.70	3	179462	4106	2923	0.9776	0.9840	0.9623	-0.03	0.01	0.04	-0.01	0.00	0.01	0.00	88.10
	Geschlecht		0.90	2	179651	4263	2734	0.9768	0.9850	0.9625	-0.03	0.01	0.03	-0.01	0.00	0.01	0.00	96.22
	Monat		0.98	7	181812	7267	573	0.9616	**0.9969**	0.9587	-0.06	0.03	0.06	0.00	-0.01	0.01	0.00	99.16
	Jahr	CLK	0.80		180580	360	1805	0.9980	0.9901	0.9882	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	97.54
	1km-Gitter		0.85		179426	95	2959	0.9995	0.9838	0.9833	0.01	0.00	-0.01	-0.01	0.00	0.00	0.00	93.82
		ECM	0.80		181077	71	1308	0.9996	0.9928	0.9924	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	97.68
			0.90		181077	71	1308	0.9996	0.9928	0.9924	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	97.68
		ECM(D)	0.80		181276	93	1109	0.9995	0.9939	**0.9934**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	98.58
			0.90		181205	89	1180	0.9995	0.9935	0.9930	0.00	0.00	0.00	0.00	0.00	0.00	0.00	98.56
20	Vorname	Exact			178883	3	3502	**1.0000**	0.9808	0.9808	0.02	0.01	-0.02	-0.01	0.00	0.01	0.00	96.31
	Nachname	MMK	0.70	2	179814	4276	2571	0.9768	0.9859	0.9633	-0.03	0.01	0.03	-0.01	0.00	0.01	0.00	97.34
	Geschlecht		0.90	2	179814	4276	2571	0.9768	0.9859	0.9633	-0.03	0.01	0.03	-0.01	0.00	0.01	0.00	97.34
	Monat		0.98	2	179814	4276	2571	0.9768	0.9859	0.9633	-0.03	0.01	0.03	-0.01	0.00	0.01	0.00	97.34
	Jahr	CLK	0.80		179961	252	2424	0.9986	0.9867	0.9853	0.01	0.00	-0.01	-0.01	0.00	0.00	0.00	97.62
			0.85		179652	107	2733	0.9994	0.9850	0.9844	0.01	0.00	-0.01	-0.01	0.00	0.00	0.00	97.23
		ECM	0.80		180651	1922	1734	0.9895	0.9905	0.9802	-0.01	0.00	0.02	0.00	0.00	0.00	0.00	98.42
			0.90		180504	1921	1881	0.9895	0.9897	0.9794	-0.01	0.00	0.02	0.00	0.00	0.01	0.00	98.18
		ECM(D)	0.80		181118	111	1267	0.9994	**0.9931**	**0.9924**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	98.80
			0.90		181118	111	1267	0.9994	**0.9931**	**0.9924**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	98.80
21	Sdx(Vorname)	Exact			160272	0	22113	**1.0000**	0.8788	0.8788	0.04	0.03	-0.04	-0.04	-0.02	0.05	0.00	0.00
	Sdx(Nachname)	MMK	0.70	2	160376	907	22009	0.9944	0.8793	0.8750	0.03	0.02	-0.02	-0.04	-0.03	0.05	0.00	0.04
	Geschlecht		0.90	6	170007	13491	12378	0.9265	0.9321	0.8679	-0.1	0.04	0.09	-0.02	-0.01	0.02	0.00	26.59
	Staatsbürgers.		0.98	9	168205	13502	14180	0.9257	0.9223	0.8587	-0.09	0.04	0.09	-0.01	-0.01	0.02	0.00	15.80
	Tag, Monat	CLK	0.80		166787	856	15598	0.9949	0.9145	0.9102	0.01	0.01	-0.02	-0.03	0.00	0.02	0.00	7.95
	Jahr		0.85		166480	125	15905	0.9992	0.9128	0.9122	0.01	0.01	-0.02	-0.03	0.00	0.02	0.00	7.69
	Postleitzahl	DTV		1	179817	3	2568	1.0000	0.9859	0.9859	0.01	0.00	-0.01	-0.01	0.00	0.00	0.00	97.16
	Straße	ECM	0.80		181442	85	943	0.9995	**0.9948**	**0.9944**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	95.11
	Hausnummer		0.90		181391	73	994	0.9996	0.9945	0.9942	0.00	0.00	0.00	0.00	0.00	0.00	0.00	95.10
22	Sdx(Vorname)	Exact			165006	0	17379	**1.0000**	0.9047	0.9047	0.04	0.02	-0.04	-0.04	-0.03	0.05	0.00	26.07
	Sdx(Nachname)	MMK	0.70	2	175474	4259	6911	0.9763	0.9621	0.9402	-0.02	0.01	0.03	0.00	-0.03	0.03	0.00	87.16
	Geschlecht		0.90	1	165006	0	17379	**1.0000**	0.9047	0.9047	0.04	0.02	-0.04	-0.04	-0.03	0.05	0.00	26.07
	Staatsbürgers.		0.98	1	165006	0	17379	**1.0000**	0.9047	0.9047	0.04	0.02	-0.04	-0.04	-0.03	0.05	0.00	26.07
	Tag, Monat	CLK	0.80		179274	945	3111	0.9948	0.9829	0.9779	0.00	0.00	0.00	-0.01	-0.01	0.01	0.00	88.08
	Jahr		0.85		175129	123	7256	0.9993	0.9602	0.9596	0.02	0.01	-0.02	-0.02	-0.01	0.02	0.00	70.19
	Postleitzahl	DTV		1	179665	3	2720	1.0000	0.9851	0.9851	0.01	0.00	-0.01	-0.01	0.00	0.00	0.00	97.16
		ECM	0.80		181423	45	962	0.9998	**0.9947**	**0.9945**	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	97.59
			0.90		181302	42	1083	0.9998	0.9941	0.9938	0.00	0.00	0.00	-0.01	0.00	0.01	0.00	97.58

Tabelle A.1.: Fortsetzung der Linkage-Ergebnisse aller getesteten Szenarien und Verfahren.

Sz.	QIDs	Methode	Schw.	n_k	TP	FP	FN	Prec.	Recall	F^*	d_{Alter}	$V_{Sf.}$	h_{Quelle}	$h_{Mig.}$	$h_{Mh.}$	$h_{Deu.}$	$h_{Ges.}$	um.(%)
23	Sdx(Vorname)	Exact			161209	0	21176	**1.0000**	0.8839	0.8839	0.03	0.02	-0.04	-0.04	-0.03	0.05	0.00	3.19
	Sdx(Nachname)	MMK	0.70	4	178720	6199	3665	0.9665	0.9799	0.9477	-0.05	0.02	0.05	0.00	0.00	0.00	0.00	81.77
	Geschlecht		0.90	7	166100	10367	16285	0.9413	0.9107	0.8617	-0.07	0.03	0.06	-0.02	-0.01	0.02	0.00	3.47
	Staatsbürgers.		0.98	3	165154	1212	17231	0.9927	0.9055	0.8995	0.00	0.01	-0.02	-0.03	0.00	0.02	0.00	3.40
	Tag, Monat	CLK	0.80		181333	10548	1052	0.9450	0.9942	0.9399	-0.08	0.04	0.10	-0.01	-0.02	0.03	-0.01	97.66
	Jahr		0.85		180230	1493	2155	0.9918	0.9882	0.9802	0.00	0.00	0.01	-0.01	-0.01	0.01	0.00	95.34
	100m-Gitter	DTV		1	179665	3	2720	1.0000	0.9851	0.9851	0.01	0.00	-0.01	-0.01	0.00	0.00	0.00	97.16
		ECM	0.80		181433	62	952	0.9997	**0.9948**	**0.9944**	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	97.16
			0.90		181430	61	955	0.9997	0.9948	0.9944	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	97.16
24	Sdx(Vorname)	Exact			163022	0	19363	**1.0000**	0.8938	0.8938	0.04	0.02	-0.04	-0.04	-0.03	0.05	0.00	14.12
	Sdx(Nachname)	MMK	0.70	3	177456	6486	4929	0.9647	0.9730	0.9396	-0.05	0.02	0.05	0.00	-0.01	0.01	0.00	81.14
	Geschlecht		0.90	1	163022	0	19363	**1.0000**	0.8938	0.8938	0.04	0.02	-0.04	-0.04	-0.03	0.05	0.00	14.12
	Staatsbürgers.		0.98	1	163022	0	19363	**1.0000**	0.8938	0.8938	0.04	0.02	-0.04	-0.04	-0.03	0.05	0.00	14.12
	Tag, Monat	CLK	0.80		181313	8978	1072	0.9528	0.9941	0.9475	-0.06	0.04	0.08	-0.01	-0.02	0.03	0.00	97.50
	Jahr		0.85		179948	1184	2437	0.9935	0.9866	0.9803	0.00	0.00	0.01	-0.01	-0.01	0.01	0.00	94.06
	1km-Gitter	DTV		1	179665	3	2720	1.0000	0.9851	0.9851	0.01	0.00	-0.01	-0.01	0.00	0.00	0.00	97.16
		ECM	0.80		181404	61	981	0.9997	**0.9946**	**0.9943**	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	97.31
			0.90		181396	60	989	0.9997	0.9946	0.9943	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	97.31
25	Sdx(Vorname)	Exact			176435	2	5950	**1.0000**	0.9674	0.9674	0.02	0.01	-0.02	-0.01	-0.03	0.03	0.00	94.58
	Sdx(Nachname)	MMK	0.70	2	179437	3834	2948	0.9791	0.9838	0.9636	-0.03	0.01	0.03	-0.01	-0.01	0.01	0.00	97.11
	Geschlecht		0.90	1	176435	2	5950	**1.0000**	0.9674	0.9674	0.02	0.01	-0.02	-0.01	-0.03	0.03	0.00	94.58
	Staatsbürgers.		0.98	5	180197	12983	2188	0.9328	0.9880	0.9223	-0.1	0.04	0.10	0.01	-0.01	0.00	0.00	97.96
	Tag, Monat	CLK	0.80		180233	3206	2152	0.9825	**0.9882**	0.9711	-0.02	0.01	0.03	-0.01	-0.02	0.02	0.00	98.19
	Jahr		0.85		178966	630	3419	0.9965	0.9813	0.9779	0.01	0.00	0.00	-0.01	-0.02	0.02	0.00	97.01
		DTV		2	179651	3	2734	1.0000	0.9850	0.9850	0.01	0.00	-0.01	-0.01	0.00	0.00	0.00	97.16
		ECM	0.80		179761	3	2624	1.0000	0.9856	**0.9856**	0.01	0.00	-0.01	-0.01	0.00	0.00	0.00	97.40
			0.90		179761	3	2624	1.0000	0.9856	**0.9856**	0.01	0.00	-0.01	-0.01	0.00	0.00	0.00	97.40
26	Sdx(Vorname)	Exact			163012	0	19373	**1.0000**	0.8938	0.8938	0.03	0.02	-0.04	-0.04	0.00	0.02	0.00	0.00
	Sdx(Nachname)	MMK	0.70	10	167122	22605	15263	0.8809	0.9163	0.8153	-0.15	0.06	0.15	-0.01	-0.02	0.02	0.00	9.80
	Geschlecht		0.90	8	168023	9482	14362	0.9466	0.9213	0.8757	-0.07	0.03	0.07	-0.02	-0.01	0.02	0.00	16.18
	Tag, Monat		0.98	10	166745	3347	15640	0.9803	0.9142	0.8978	-0.02	0.01	0.01	-0.03	0.00	0.02	0.00	8.59
	Jahr	CLK	0.80		166794	965	15591	0.9942	0.9145	0.9097	0.01	0.01	-0.02	-0.03	0.00	0.02	0.00	7.92
	Postleitzahl		0.85		166483	134	15902	0.9992	0.9128	0.9121	0.01	0.01	-0.02	-0.03	0.00	0.02	0.00	7.66
	Straße	DTV		1	179817	3	2568	1.0000	0.9859	0.9859	0.01	0.00	-0.01	-0.01	0.00	0.00	0.00	97.16
	Hausnummer	ECM	0.80		181754	85	631	0.9995	**0.9965**	**0.9961**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	97.02
			0.90		170241	84	12144	0.9995	0.9334	0.9330	0.01	0.01	-0.02	-0.03	0.00	0.02	0.00	27.82
27	Sdx(Vorname)	Exact			167890	0	14495	**1.0000**	0.9205	0.9205	0.03	0.02	-0.04	-0.04	0.00	0.02	0.00	26.69
	Sdx(Nachname)	MMK	0.70	1	168091	453	14294	0.9973	0.9216	0.9193	0.02	0.02	-0.03	-0.03	0.00	0.02	0.00	27.85
	Geschlecht		0.90	1	168091	453	14294	0.9973	0.9216	0.9193	0.02	0.02	-0.03	-0.03	0.00	0.02	0.00	27.85
	Tag, Monat		0.98	1	168091	453	14294	0.9973	0.9216	0.9193	0.02	0.02	-0.03	-0.03	0.00	0.02	0.00	27.85
	Jahr	CLK	0.80		178985	712	3400	0.9960	0.9814	0.9775	0.00	0.00	0.00	-0.01	0.00	0.01	0.00	84.30
	Postleitzahl		0.85		175762	121	6623	0.9993	0.9637	0.9630	0.01	0.01	-0.02	-0.02	0.00	0.01	0.00	66.52
		DTV		1	179665	3	2720	1.0000	0.9851	0.9851	0.01	0.00	-0.01	-0.01	0.00	0.00	0.00	97.16
		ECM	0.80		181475	50	910	0.9997	**0.9950**	**0.9947**	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	97.61
			0.90		181373	50	1012	0.9997	0.9945	0.9942	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	97.36
28	Sdx(Vorname)	Exact			164015	0	18370	**1.0000**	0.8993	0.8993	0.02	0.02	-0.04	-0.04	0.00	0.02	0.00	3.33
	Sdx(Nachname)	MMK	0.70	3	164921	1224	17464	0.9926	0.9042	0.8982	0.01	0.01	-0.02	-0.03	0.00	0.02	0.00	3.73
	Geschlecht		0.90	3	164568	489	17817	0.9970	0.9023	0.8999	0.02	0.02	-0.03	-0.04	0.00	0.02	0.00	4.82
	Tag, Monat		0.98	1	164266	424	18119	0.9974	0.9007	0.8986	0.02	0.02	-0.03	-0.04	0.00	0.02	0.00	4.79
	Jahr	CLK	0.80		181460	9878	925	0.9484	**0.9949**	0.9438	-0.07	0.04	0.08	0.00	-0.01	0.01	0.00	98.20
	100m-Gitter		0.85		180725	1250	1660	0.9931	0.9909	0.9842	0.00	0.00	0.01	0.00	0.00	0.00	0.00	95.22
		DTV		1	179665	3	2720	1.0000	0.9851	0.9851	0.01	0.00	-0.01	-0.01	0.00	0.00	0.00	97.16
		ECM	0.80		181432	62	953	0.9997	0.9948	**0.9944**	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	97.16
			0.90		181430	62	955	0.9997	0.9948	0.9944	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	97.16

Tabelle A.1.: Fortsetzung der Linkage-Ergebnisse aller getesteten Szenarien und Verfahren.

Sz.	QIDs	Methode	Schw.	n_k	TP	FP	FN	Prec.	Recall	F^*	d_{Alter}	$V_{Sf.}$	h_{Quelle}	$h_{Mig.}$	$h_{Mh.}$	$h_{Deu.}$	$h_{Ges.}$	um.(%)
29	Sdx(Vorname)	Exact			165877	0	16508	**1.0000**	0.9095	0.9095	0.03	0.02	-0.04	-0.04	0.00	0.02	0.00	14.56
	Sdx(Nachname)	MMK	0.70	1	166104	438	16281	0.9974	0.9107	0.9086	0.02	0.02	-0.03	-0.03	0.00	0.02	0.00	15.87
	Geschlecht		0.90	1	166104	438	16281	0.9974	0.9107	0.9086	0.02	0.02	-0.03	-0.03	0.00	0.02	0.00	15.87
	Tag, Monat		0.98	1	166104	438	16281	0.9974	0.9107	0.9086	0.02	0.02	-0.03	-0.03	0.00	0.02	0.00	15.87
	Jahr	CLK	0.80		181421	8276	964	0.9564	0.9947	0.9515	-0.06	0.03	0.07	0.00	-0.01	0.01	0.00	97.92
	1km-Gitter		0.85		180385	1007	2000	0.9944	0.9890	0.9836	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	93.26
		DTV		1	179665	3	2720	1.0000	0.9851	0.9851	0.01	0.00	-0.01	-0.01	0.00	0.00	0.00	97.16
		ECM	0.80		181451	63	934	0.9997	**0.9949**	0.9945	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	97.32
			0.90		181434	39	951	0.9998	0.9948	**0.9946**	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	97.32
30	Sdx(Vorname)	Exact			179647	3	2738	1.0000	0.9850	0.9850	0.01	0.00	-0.01	-0.01	0.00	0.00	0.00	97.16
	Sdx(Nachname)	MMK	0.70	3	179913	3551	2472	0.9806	0.9864	0.9676	-0.02	0.01	0.03	0.00	0.00	0.01	0.00	97.52
	Geschlecht		0.90	1	179657	541	2728	0.9970	0.9850	0.9821	0.01	0.00	-0.01	-0.01	0.00	0.00	0.00	97.16
	Tag, Monat		0.98	1	179657	541	2728	0.9970	0.9850	0.9821	0.01	0.00	-0.01	-0.01	0.00	0.00	0.00	97.16
	Jahr	CLK	0.80		181178	2466	1207	0.9866	**0.9934**	0.9801	-0.02	0.01	0.02	0.00	0.00	0.00	0.00	98.94
			0.85		180857	539	1528	0.9970	0.9916	**0.9887**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	98.60
		DTV		1	179647	3	2738	1.0000	0.9850	0.9850	0.01	0.00	-0.01	-0.01	0.00	0.00	0.00	97.16
		ECM	0.80		179786	3	2599	**1.0000**	0.9857	0.9857	0.01	0.00	-0.01	-0.01	0.00	0.00	0.00	97.40
			0.90		179786	3	2599	**1.0000**	0.9857	0.9857	0.01	0.00	-0.01	-0.01	0.00	0.00	0.00	97.40
31	Sdx(Vorname)	Exact			160326	0	22059	**1.0000**	0.8791	0.8791	0.04	0.03	-0.04	-0.04	-0.02	0.05	0.00	0.00
	Sdx(Nachname)	MMK	0.70	2	160424	907	21961	0.9944	0.8796	0.8752	0.03	0.02	-0.02	-0.04	-0.03	0.05	0.00	0.04
	Geschlecht		0.90	2	160424	907	21961	0.9944	0.8796	0.8752	0.03	0.02	-0.02	-0.04	-0.03	0.05	0.00	0.04
	Staatsbürgers.		0.98	7	167197	10295	15188	0.9420	0.9167	0.8677	-0.08	0.03	0.07	-0.02	-0.01	0.02	0.00	12.90
	Monat	CLK	0.80		166739	2320	15646	0.9863	0.9142	0.9027	0.00	0.01	-0.01	-0.03	0.00	0.02	0.00	7.89
	Jahr		0.85		166396	376	15989	0.9977	0.9123	0.9105	0.01	0.01	-0.02	-0.03	0.00	0.02	0.00	7.54
	Postleitzahl	ECM	0.80		170077	172	12308	0.9990	**0.9325**	**0.9316**	0.01	0.01	-0.02	-0.03	0.00	0.02	0.00	26.95
	Straße		0.90		170072	167	12313	0.9990	0.9325	0.9316	0.01	0.01	-0.02	-0.03	0.00	0.02	0.00	26.94
	Hausnummer																	
32	Sdx(Vorname)	Exact			165063	0	17322	**1.0000**	0.9050	0.9050	0.04	0.02	-0.04	-0.04	-0.03	0.05	0.00	26.10
	Sdx(Nachname)	MMK	0.70	3	176557	8324	5828	0.9550	0.9680	0.9258	-0.06	0.03	0.06	0.00	-0.01	0.01	0.00	77.65
	Geschlecht		0.90	2	165810	5877	16575	0.9658	0.9091	0.8807	-0.02	0.01	0.03	-0.04	-0.04	0.06	0.00	26.34
	Staatsbürgers.		0.98	1	165063	0	17322	**1.0000**	0.9050	0.9050	0.04	0.02	-0.04	-0.04	-0.03	0.05	0.00	26.10
	Monat	CLK	0.80		178159	4010	4226	0.9780	0.9768	0.9558	-0.02	0.01	0.03	-0.01	-0.01	0.02	0.00	82.38
	Jahr		0.85		174228	755	8157	0.9957	0.9553	0.9513	0.01	0.01	-0.01	-0.02	-0.02	0.03	0.00	65.55
	Postleitzahl	ECM	0.80		180342	81	2043	0.9996	**0.9888**	**0.9884**	0.01	0.00	-0.01	-0.01	0.00	0.01	0.00	95.30
			0.90		179430	25	2955	0.9999	0.9838	0.9837	0.01	0.00	-0.01	-0.01	0.00	0.01	0.00	95.02
33	Sdx(Vorname)	Exact			161265	0	21120	**1.0000**	0.8842	0.8842	0.03	0.02	-0.04	-0.04	-0.03	0.05	0.00	3.19
	Sdx(Nachname)	MMK	0.70	6	180771	10863	1614	0.9433	0.9912	0.9354	-0.09	0.04	0.09	0.01	0.00	0.00	0.00	92.85
	Geschlecht		0.90	3	165466	6173	16919	0.9640	0.9072	0.8775	-0.04	0.01	0.03	-0.03	-0.01	0.03	0.00	3.42
	Staatsbürgers.		0.98	7	180843	6506	1542	0.9653	0.9915	0.9574	-0.05	0.02	0.05	0.01	0.00	0.00	0.00	97.68
	Monat	CLK	0.80		180102	18972	2283	0.9047	0.9875	0.8944	-0.13	0.07	0.16	-0.01	-0.04	0.04	-0.01	92.17
	Jahr		0.85		179092	6417	3293	0.9654	0.9819	0.9486	-0.04	0.02	0.05	-0.01	-0.02	0.02	0.00	88.27
	100m-Gitter	ECM	0.80		180984	82	1401	0.9995	**0.9923**	**0.9919**	0.00	0.00	0.00	-0.01	0.00	0.01	0.00	94.55
			0.90		180984	82	1401	0.9995	**0.9923**	**0.9919**	0.00	0.00	0.00	-0.01	0.00	0.01	0.00	94.55
34	Sdx(Vorname)	Exact			163077	0	19308	**1.0000**	0.8941	0.8941	0.04	0.02	-0.04	-0.04	-0.03	0.05	0.00	14.13
	Sdx(Nachname)	MMK	0.70	4	176363	9629	6022	0.9482	0.9670	0.9185	-0.07	0.03	0.07	0.00	-0.01	0.01	0.00	73.47
	Geschlecht		0.90	2	163929	3789	18456	0.9774	0.8988	0.8805	-0.01	0.01	0.01	-0.03	-0.03	0.05	0.00	14.24
	Staatsbürgers.		0.98	1	163077	0	19308	**1.0000**	0.8941	0.8941	0.04	0.02	-0.04	-0.04	-0.03	0.05	0.00	14.13
	Monat	CLK	0.80		180214	18597	2171	0.9065	0.9881	0.8967	-0.12	0.06	0.15	-0.02	-0.04	0.04	0.00	92.94
	Jahr		0.85		179034	5826	3351	0.9685	0.9816	0.9512	-0.03	0.02	0.05	-0.01	-0.02	0.02	0.00	88.60
	1km-Gitter	ECM	0.80		180348	67	2037	0.9996	**0.9888**	**0.9885**	0.01	0.00	-0.01	-0.01	0.00	0.01	0.00	94.93
			0.90		180283	66	2102	0.9996	0.9885	0.9881	0.01	0.00	-0.01	-0.01	0.00	0.01	0.00	94.92

Tabelle A.1.: Fortsetzung der Linkage-Ergebnisse aller getesteten Szenarien und Verfahren.

Sz.	QIDs	Methode	Schw.	n_k	TP	FP	FN	Prec.	Recall	F^*	d_{Alter}	$V_{Sf.}$	h_{Quelle}	$h_{Mig.}$	$h_{Mh.}$	$h_{Deu.}$	$h_{Ges.}$	um.(%)
35	Sdx(Vorname)	Exact			176214	26	6171	**0.9999**	0.9662	0.9660	0.02	0.01	-0.01	-0.01	-0.03	0.03	0.00	94.54
	Sdx(Nachname)	MMK	0.70	2	179153	3838	3232	0.9790	0.9823	0.9620	-0.03	0.01	0.03	-0.01	-0.01	0.01	0.00	96.97
	Geschlecht		0.90	2	176500	4364	5885	0.9759	0.9677	0.9451	-0.02	0.01	0.03	0.00	-0.03	0.03	0.00	94.89
	Staatsbürgers.		0.98	3	177446	5066	4939	0.9722	0.9729	0.9466	-0.03	0.01	0.04	0.01	-0.02	0.01	0.00	95.71
	Monat	CLK	0.80		179147	10569	3238	0.9443	0.9822	0.9284	-0.07	0.04	0.10	-0.02	-0.03	0.04	0.00	97.37
	Jahr		0.85		178368	3589	4017	0.9803	0.9780	0.9591	-0.02	0.01	0.03	-0.01	-0.03	0.03	0.00	96.59
		ECM	0.80		179405	33	2980	0.9998	**0.9837**	**0.9835**	0.01	0.00	-0.01	-0.01	0.00	0.00	0.00	97.11
			0.90		179405	33	2980	0.9998	**0.9837**	**0.9835**	0.01	0.00	-0.01	-0.01	0.00	0.00	0.00	97.11
36	Sdx(Vorname)	Exact			163069	0	19316	**1.0000**	0.8941	0.8941	0.03	0.02	-0.04	-0.04	0.00	0.02	0.00	0.00
	Sdx(Nachname)	MMK	0.70	2	163152	918	19233	0.9944	0.8945	0.8901	0.02	0.02	-0.02	-0.04	0.00	0.03	0.00	0.04
	Geschlecht		0.90	4	167649	12562	14736	0.9303	0.9192	0.8600	-0.09	0.04	0.09	-0.02	-0.01	0.03	0.00	15.29
	Monat		0.98	9	169778	9867	12607	0.9451	0.9309	0.8831	-0.07	0.03	0.07	-0.02	-0.01	0.02	0.00	26.57
	Jahr	CLK	0.80		166742	2467	15643	0.9854	0.9142	0.9020	0.00	0.01	-0.01	-0.02	0.00	0.01	0.00	7.87
	Postleitzahl		0.85		166432	397	15953	0.9976	0.9125	0.9105	0.01	0.01	-0.02	-0.03	0.00	0.02	0.00	7.54
	Straße	ECM	0.80		170159	326	12226	0.9981	**0.9330**	**0.9313**	0.01	0.01	-0.02	-0.03	0.00	0.02	0.00	27.21
	Hausnummer		0.90		170159	326	12226	0.9981	**0.9330**	**0.9313**	0.01	0.01	-0.02	-0.03	0.00	0.02	0.00	27.21
37	Sdx(Vorname)	Exact			167946	0	14439	**1.0000**	0.9208	0.9208	0.03	0.02	-0.03	-0.04	0.00	0.02	0.00	26.72
	Sdx(Nachname)	MMK	0.70	3	177917	3392	4468	0.9813	0.9755	0.9577	-0.02	0.01	0.02	-0.01	0.00	0.01	0.00	85.49
	Geschlecht		0.90	1	168147	453	14238	0.9973	0.9219	0.9197	0.02	0.02	-0.03	-0.03	0.00	0.02	0.00	27.88
	Monat		0.98	1	168147	453	14238	0.9973	0.9219	0.9197	0.02	0.02	-0.03	-0.03	0.00	0.02	0.00	27.88
	Jahr	CLK	0.80		177585	3462	4800	0.9809	0.9737	0.9555	-0.01	0.00	0.02	-0.01	0.00	0.01	0.00	76.24
	Postleitzahl		0.85		174696	662	7689	0.9962	0.9578	0.9544	0.01	0.01	-0.02	-0.02	0.00	0.01	0.00	60.35
		ECM	0.80		179887	34	2498	0.9998	**0.9863**	**0.9861**	0.01	0.00	-0.01	-0.01	0.00	0.00	0.00	96.99
			0.90		179783	32	2602	0.9998	0.9857	0.9856	0.01	0.00	-0.01	-0.01	0.00	0.00	0.00	96.96
38	Sdx(Vorname)	Exact			164074	0	18311	**1.0000**	0.8996	0.8996	0.02	0.02	-0.04	-0.04	0.00	0.02	0.00	3.33
	Sdx(Nachname)	MMK	0.70	3	164927	1224	17458	0.9926	0.9043	0.8983	0.01	0.01	-0.02	-0.03	0.00	0.02	0.00	3.73
	Geschlecht		0.90	3	164927	1224	17458	0.9926	0.9043	0.8983	0.01	0.01	-0.02	-0.03	0.00	0.02	0.00	3.73
	Monat		0.98	3	164570	489	17815	0.9970	0.9023	0.8999	0.02	0.02	-0.03	-0.04	0.00	0.02	0.00	4.82
	Jahr	CLK	0.80		180040	21286	2345	0.8943	0.9871	0.8840	-0.14	0.07	0.16	0.01	-0.02	0.01	-0.01	90.62
	100m-Gitter		0.85		179147	6705	3238	0.9639	0.9822	0.9474	-0.04	0.02	0.05	0.00	-0.01	0.01	0.00	85.94
		ECM	0.80		181438	90	947	0.9995	**0.9948**	0.9943	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	96.97
			0.90		181421	66	964	0.9996	0.9947	**0.9944**	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	96.97
39	Sdx(Vorname)	Exact			165933	0	16452	**1.0000**	0.9098	0.9098	0.03	0.02	-0.04	-0.04	0.00	0.02	0.00	14.56
	Sdx(Nachname)	MMK	0.70	3	166396	627	15989	0.9962	0.9123	0.9092	0.02	0.02	-0.03	-0.03	0.00	0.02	0.00	15.87
	Geschlecht		0.90	1	166160	438	16225	0.9974	0.9110	0.9089	0.02	0.02	-0.03	-0.03	0.00	0.02	0.00	15.88
	Monat		0.98	1	166160	438	16225	0.9974	0.9110	0.9089	0.02	0.02	-0.03	-0.03	0.00	0.02	0.00	15.88
	Jahr	CLK	0.80		180156	20388	2229	0.8983	0.9878	0.8885	-0.13	0.07	0.15	0.01	-0.02	0.01	0.00	91.34
	1km-Gitter		0.85		179059	5870	3326	0.9683	0.9818	0.9512	-0.03	0.02	0.04	-0.01	-0.01	0.01	0.00	85.56
		ECM	0.80		180775	77	1610	0.9996	**0.9912**	**0.9908**	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	97.06
			0.90		180775	77	1610	0.9996	**0.9912**	**0.9908**	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	97.06
40	Sdx(Vorname)	Exact			179265	32	3120	**0.9998**	0.9829	**0.9827**	0.01	0.00	-0.01	-0.01	0.00	0.01	0.00	97.02
	Sdx(Nachname)	MMK	0.70	2	179477	3887	2908	0.9788	0.9841	0.9635	-0.03	0.01	0.03	0.00	0.00	0.01	0.00	97.28
	Geschlecht		0.90	2	179481	4001	2904	0.9782	0.9841	0.9630	-0.03	0.01	0.03	0.00	0.00	0.01	0.00	97.29
	Monat		0.98	1	179275	570	3110	0.9968	0.9829	0.9799	0.01	0.00	-0.01	-0.01	0.00	0.00	0.00	97.02
	Jahr	CLK	0.80		180530	9454	1855	0.9502	**0.9898**	0.9410	-0.06	0.03	0.08	0.00	-0.01	0.01	0.00	98.47
			0.85		180297	3430	2088	0.9813	0.9886	0.9703	-0.02	0.01	0.03	0.00	0.00	0.01	0.00	98.22
		ECM	0.80		179265	32	3120	**0.9998**	0.9829	**0.9827**	0.01	0.00	-0.01	-0.01	0.00	0.01	0.00	97.02
			0.90		179265	32	3120	**0.9998**	0.9829	**0.9827**	0.01	0.00	-0.01	-0.01	0.00	0.01	0.00	97.02

	Szenario A		Szenario B		Szenario C	
Variable	Logit	SE_{Logit}	Logit	SE_{Logit}	Logit	SE_{Logit}
Konstante	5.255^{***}	0.058	6.457^{***}	0.095	6.756^{***}	0.118
Weiblich	0.083^{*}	0.047	0.123^{*}	0.072	0.105	0.088
Herkunft						
Deutsch						
Migrationshintergrund	-0.369^{***}	0.065	-0.633^{***}	0.114	-0.001	0.157
Migrant	-1.207^{***}	0.055	-2.174^{***}	0.083	-1.735^{***}	0.097
Schulform						
Grundschule						
Stadtteilschule	-0.441^{***}	0.063	-0.271^{***}	0.096	-0.418^{***}	0.128
Gymnasium	0.068	0.079	0.171	0.128	0.030	0.162
Berufsschule	-0.856^{***}	0.101	-0.819^{***}	0.149	-1.036^{***}	0.184
Andere	-0.512^{***}	0.140	-0.263	0.225	-0.563^{**}	0.267
Wohnort (RISE)						
Hoher Status (3-4)						
Niedriger Status (1-2)	-0.018	0.056	-0.005	0.083	0.153	0.110
Außerhalb Hamburgs	0.016	0.103	-0.395^{***}	0.150	-0.371^{**}	0.160
Schuldaten	-0.838^{***}	0.096	-0.664^{***}	0.143	-0.879^{***}	0.168
Log Pseudo-Likelihood	$-9,860.953$		$-4,578.595$		$-3,246.700$	
χ^2	1,229.640		1,104.931		721.911	
p	0.000		0.000		0.000	
R^2 (Nagelkerk)	0.062		0.110		0.102	
n	182,406		182,444		182,461	
FP+FN	1,881		798		527	

$^{*}p < 0,05$ $^{**}p < 0,01$ $^{***}p < 0,001$

Tabelle A.2.: Logits und Standardfehler der Logits (SE_{Logit}) des logistischen Regressionsmodells für das unmodifizierte probabilistische Record-Linkage. Die abhängige Variable ist die Gültigkeit der Verknüpfung (0 = falsch verknüpft, 1 = richtig verknüpft).

	Szenario A		Szenario B		Szenario C	
Variable	Logit	SE_{Logit}	Logit	SE_{Logit}	Logit	SE_{Logit}
Konstante	5.465^{***}	0.066	6.911^{***}	0.126	6.344^{***}	0.092
Weiblich	0.078	0.055	0.149	0.094	-0.002	0.069
Herkunft						
Deutsch						
Migrationshintergrund	-0.246^{***}	0.080	-0.024	0.170	-0.229^{**}	0.112
Migrant	-1.215^{***}	0.064	-1.776^{***}	0.102	-1.608^{***}	0.076
Schulform						
Grundschule						
Stadtteilschule	-0.254^{***}	0.073	-0.382^{***}	0.137	-0.364^{***}	0.099
Gymnasium	0.338^{***}	0.096	0.127	0.179	0.229^{*}	0.133
Berufsschule	-0.530^{***}	0.119	-1.257^{***}	0.192	-1.090^{***}	0.143
Andere	-0.321^{*}	0.164	-0.271	0.324	-0.630^{***}	0.200
Wohnort (RISE)						
Hoher Status (3-4)						
Niedriger Status (1-2)	-0.081	0.065	0.049	0.113	-0.169^{**}	0.079
Außerhalb Hamburgs	0.060	0.118	-0.284^{*}	0.168	-0.082	0.132
Schuldaten	-1.081^{***}	0.112	-0.805^{***}	0.175	-1.009^{***}	0.129
Log Pseudo-Likelihood	$-7,513.110$		$-2,909.144$		$-4,948.853$	
χ^2	969.432		736.359		1,194.690	
p	0.000		0.000		0.000	
R^2 (Nagelkerk)	0.063		0.114		0.111	
n	182,430		182,476		182,615	
FP+FN	1,356		471		875	

$^{*}p < 0,05$ $^{**}p < 0,01$ $^{***}p < 0,001$

Tabelle A.3.: Logits und Standardfehler der Logits (SE_{Logit}) des logistischen Regressionsmodells für das modifizierte probabilistische Record-Linkage. Die abhängige Variable ist die Gültigkeit der Verknüpfung (0 = falsch verknüpft, 1 = richtig verknüpft).

	Szenario A		Szenario B		Szenario C	
Variable	Logit	SE_{Logit}	Logit	SE_{Logit}	Logit	SE_{Logit}
Konstante	5.174***	0.056	5.448***	0.063	2.544***	0.018
Weiblich	0.123***	0.042	0.069	0.048	−0.070***	0.017
Herkunft						
Deutsch						
Migrationshintergrund	−0.193***	0.059	−0.250***	0.069	−0.181***	0.022
Migrant	−0.880***	0.050	−1.051***	0.057	−0.867***	0.022
Schulform						
Grundschule						
Stadtteilschule	−0.815***	0.060	−0.619***	0.069	0.090***	0.021
Gymnasium	−0.277***	0.072	−0.128	0.083	0.320***	0.024
Berufsschule	−1.229***	0.092	−1.136***	0.103	−0.403***	0.040
Andere	−0.870***	0.123	−0.791***	0.140	−0.112**	0.052
Wohnort (RISE)						
Hoher Status (3-4)						
Niedriger Status (1-2)	−0.056	0.050	−0.059	0.058	−0.002	0.021
Außerhalb Hamburgs	0.252***	0.090	−0.098	0.096	−0.309***	0.046
Schuldaten	−0.968***	0.082	−0.874***	0.095	−0.302***	0.043
Log Pseudo-Likelihood	−11,999.946		−9,393.410		−52,885.997	
χ^2	1,678.448		1,332.344		3,024.983	
p	0.000		0.000		0.000	
R^2 (Nagelkerk)	0.070		0.070		0.037	
n	182,465		182,548		182,441	
FP+FN	2,414		1,790		16,077	

*$p < 0,05$ **$p < 0,01$ ***$p < 0,001$

Tabelle A.4.: Logits und Standardfehler der Logits (SE_{Logit}) des logistischen Regressionsmodells für Record-Linkage mit CLKs und Multibit-Trees. Die abhängige Variable ist die Gültigkeit der Verknüpfung (0 = falsch verknüpft, 1 = richtig verknüpft).

	Szenario A		Szenario B		Szenario C	
Variable	Logit	SE_{Logit}	Logit	SE_{Logit}	Logit	SE_{Logit}
Konstante	5.392^{***}	0.062	6.578^{***}	0.103	6.450^{***}	0.096
Weiblich	0.036	0.052	0.173^{**}	0.078	0.174^{**}	0.074
Herkunft						
Deutsch						
Migrationshintergrund	-0.458^{***}	0.073	-0.607^{***}	0.123	-0.785^{***}	0.109
Migrant	-1.357^{***}	0.061	-2.082^{***}	0.089	-2.039^{***}	0.086
Schulform						
Grundschule						
Stadtteilschule	-0.153^{**}	0.068	-0.292^{***}	0.104	-0.241^{**}	0.098
Gymnasium	0.429^{***}	0.091	0.325^{**}	0.147	0.229^{*}	0.132
Berufsschule	-0.551^{***}	0.112	-0.677^{***}	0.162	-0.762^{***}	0.155
Andere	-0.167	0.162	-0.122	0.254	-0.322	0.221
Wohnort (RISE)						
Hoher Status (3-4)						
Niedriger Status (1-2)	-0.077	0.061	-0.029	0.090	-0.094	0.083
Außerhalb Hamburgs	-0.049	0.113	-0.427^{***}	0.152	-0.025	0.169
Schuldaten	-0.964^{***}	0.107	-0.930^{***}	0.152	-0.791^{***}	0.147
Log Pseudo-Likelihood	$-8,188.397$		$-3,999.035$		$-4,431.854$	
χ^2	1,128.054		975.802		940.183	
p	0.000		0.000		0.000	
R^2 (Nagelkerk)	0.067		0.111		0.098	
n	182,405		182,440		182,439	
FP+FN	1,512		681		756	

$^{*}p < 0,05$ $^{**}p < 0,01$ $^{***}p < 0,001$

Tabelle A.5.: Logits und Standardfehler der Logits (SE_{Logit}) des logistischen Regressionsmodells für das Record-Linkage nach dem Datavant-Verfahren. Die abhängige Variable ist die Gültigkeit der Verknüpfung (0 = falsch verknüpft, 1 = richtig verknüpft).

IMPRINT

Publisher

German Record-Linkage Center
Regensburger Str. 100
D-90478 Nuremberg

Editor

Rainer Schnell

Template layout

Christine Weidmann

Download

www.record-linkage.de

The German Record Linkage Center is funded by the German Research Foundation (DFG).

DOI: 10.17185/duepublico/81929
URN: urn:nbn:de:hbz:465-20240506-142841-8