Ludwig-Maximilians-Universität Mnchen
Institut für Statistik

Vorbereitungsmaterial zum Thema

# Record Linkage

Im Rahmen des Seminars
„Spezielle Themen der Sozial- und Wirtschaftsstatistik“

verfasst von
Alexander Pokatilo

**Betreuer:** Paul Fink

# Inhaltsverzeichnis

# 1 Einführung

In Forschungspraxis kann es oft von Interesse, sein große Datensätze mit vielen Variablen zu betrachten. Die Erschaffung solcher Datensätze kann allerdings sowohl sehr zeit-, als auch kostenaufwendig sein. Außerdem stellen umfangreiche Fragebögen eine große Herausforderung für die befragten Personen dar, was die Teilnahmebereitschaft an der Studie beeinflussen kann; manche Daten (z.B. medizinische) können auf der Personenebene nur ungenau oder gar nicht gesammelt werden. Alternativ zur Erhebung von vielen Variablen kann man versuchen, bereits vorhandene Informationen zu benutzen. So kann man beispielsweise Survey-Daten durch administrative Daten ergänzen. Die Zusammenführung von zwei oder mehreren Datensets, die aus unterschiedlichen Datenquellen stammen, aber die gleichen Objekte (Personen, Patienten, Kunden, Unternehmen) enthalten, wird auch **record linkage** genannt. Record linkage kann man auch als eine spezielle Form der Sekundäranalyse betrachten bzw. als einen Prozess der Identifikation von Datensätzen, die auf dieselbe Untersuchungseinheit in zwei oder mehreren unabhängigen Datensets verweisen. Seit der Entstehung der ersten theoretischen Vorüberlegungen am Ende der 60-er Jahre zieht das Thema immer mehr Interesse der Forscher an, was die fast kontinuierlich wachsende Anzahl von wissenschaftlichen Publikationen zum Thema bestätigt (S. Abbildung 1.1). An dieser Stelle muss man auch erwähnen, dass man die Methode des record linkage in der Literatur auch unter den Namen „record matching“, „object identification“oder „duplicate detection“finden kann. Wichtig ist es, record linkage vom *statistischen Matching* zu trennen. Dabei werden auch Datensätze aus verschiedenen Quellen zusammengeführt, die Objekte sind aber auch unterschiedlich.
Record linkage wird auch oft durchgeführt, um Duplikate in einem Datenset zu finden und zu entfernen, um die Qualität der Daten zu verbessern (sog. *Data Cleaning*). Als Beispiel für diese Problemstellung kann man sich zwei große Unternehmen vorstellen, die in der gleichen Branche tätig sind. Wird eins von diesen Unternehmen das andere übernehmen, so müssen unter anderem die Kundendaten aggregiert werden. Um eine effektive Kundenbetreuung zu gewhrleisteten sollen alle Kunden in der gemeinsamen Datenbank nur einmal vorkommen. Es ist aber sehr wahrscheinlich, dass die Daten-

sets über keine einheitliche Schreibweise von Eigennamen oder keine Formatierung von Adressdaten verfügen, was die Identifikation von Duplikaten erschwert.

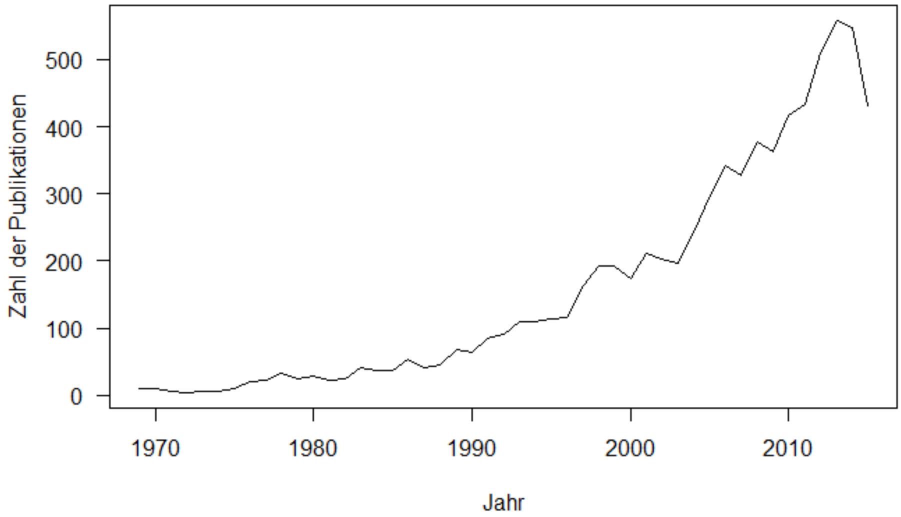


**Abbildung 1.1:** Anzahl der record linkage Anwendungen in der medizinischen Literatur (Datenbank Pubmed)

Man kann sich gut vorstellen, dass das Zusammenführen von Datensets relativ einfach verläuft, wenn die Datensätze über ein global eindeutiges Merkmal, wie z.B. Personennummer verfügen. In der Praxis ist es aber ein eher seltener Fall. So wird im deutschen medizinischen System keine allgemeine Identifizierung von Patientendaten verwendet. Im Gegensatz dazu, gibt es in den skandinavischen Ländern eine *Personennummer*, die aus dem sechsstelligen Geburtstag in der Form JJMMTT (Schweden) bzw. TTMMJJ (Dänemark, Finnland und Norwegen) gefolgt von 4 Ziffern (in Norwegen 5 Ziffern), die unter anderem das Geschlecht kodieren, besteht. Diese Nummer wird sowohl von öffentlichen als auch nicht-öffentlichen Stellen als Schlüssel verwendet.

Angesichts des fehlenden eindeutigen Schlüssels muss man auf andere identifizierende

| Vorname | Nachname | Straße | PLZ | Stadt | Geschlecht | Geburtsjahr |
|---|---|---|---|---|---|---|
| Robb | Stark | Nordstadtstr. 10 | 38448 | Wolfsburg | m | 1996 |
| Robert | Stark | Nordstr. 1 | 34848 | Wolfsburg | m | 1995 |
| R. | Stark | Nordstadtstr. 1 | 38448 | NA | m | 2000 |

**Tabelle 1.1:** Fehlerhafte Daten erschweren die Datenverknüpfung

Variablen zugreifen, wie Namen, Geburtsdatum, Geschlecht, Adresse usw. Dies ist mit zusätzlichen Herausforderungen verbunden, da diese Informationen oft von unterschiedlichen Fehlern (typographische Fehler, fehlende Werte, veraltete Daten, unterschiedliche Kodierung) behaftet sind (S. Tabelle 1.1). Besonders fehleranfällig sind Namen und Adressen, da es oft mehrere Variationen eines Namens gibt und somit unklar ist, ob ein Tippfehler vorliegt oder nicht (z.B. Stefan Meyer, Steffan Mayer, Steffen Meier, Stephan Maier, Steph Mayr), auch mögliche Veränderung des Namens durch Heirat oder der Adresse in Folge eines Wohnrotwechsels sollen berücksichtigt werden. Somit soll die Anwendung von record linkage eine gewisse Fehlertoleranz anbieten, so dass die Zusammenführung von Datensätzen auch bei nicht vollständig übereinstimmenden Identifikationsmerkmalen möglich wird (Hentschel [2008, S.62]).
Im Folgenden soll das Prozess des record linkage ausführlicher präsentiert werden. Dabei wird auf einzelne Etappen und damit verbundene Probleme eingegangen.

# 2 Theorie

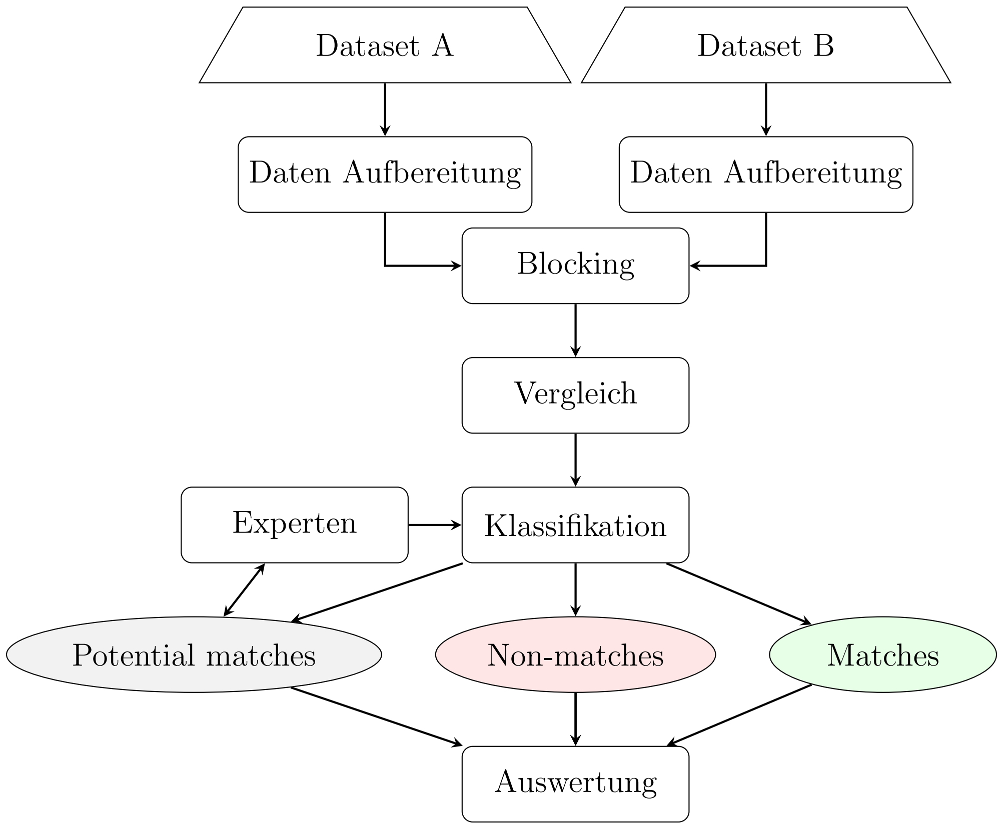


**Abbildung 2.1:** Record linkage Prozess

Der Begriff record linkage wurde zum ersten Mal vom Leiter des U.S National Office of Vital Statistics Dr. Halbert L. Dunn (Dunn [1946]) verwendet.
Ein Überblick über den Verlauf von record linkage wird in der Abbildung 2.1 gegeben. Dabei besteht dieser Prozess aus fünf wesentlichen Schritten: im ersten Schritt (*Data Pre-Processing*) werden die Daten aus unterschiedlichen Quellen zu einem einheitlichen Format gebracht bzw. aufbereitet. Dann werden die Daten in Blöcke aufgeteilt, um die Anzahl der Vergleichspaare und somit die Analysedauer zu reduzieren. Im dritten Schritt werden die Paare von Datensätzen miteinander verglichen und eine Art Ähnlich-

keitsscore gebildet. In der Klassifikationsphase werden die Paare anhand von Scores in drei Gruppen aufgeteilt, nämlich in Matches, Non-Matches und potentielle Matches. Im letzten Schritt werden die Paare, die als Matches klassifiziert wurden, zu einem Master-Datensatzverknüpft. In diesem Schritt wird auch die Vollständigkeit und die Qualität der gematchten Datensätze beurteilt.

## 2.1 Aufbereitung der Daten

Da die Datensets beim record linkage aus unterschiedlichen Quellen stammen, können sie auch unterschiedliche Form und Struktur bei einzelnen Variablen aufweisen (S. Tabelle 2.1). Da der Linkage-Prozess auf Basis von persönlichen Daten erfolgt, sollen diese Informationen im ersten Schritt bereinigt und standardisiert werden. Zweck der Standardisierung besteht darin, Variationen in den verschiedenen Variablenwerten gering zu halten und möglichst viele Fehler bereits vor dem eigentlichen Linkage zu beseitigen (Nasseh [2014, S.18]). Dieser Vorbereitungsschritt ist essenziell für ein erfolgreiches record linkage.
Die Datenaufbereitungsphase kann je nach Datenset aus folgenden Etappen bestehen:

- Entfernung von ungeeigneten Zeichen und Wörter, Ersetzung von undeutschen Sonderzeichen (z.B. č → c)
- Ersetzung von Abkürzungen und Korrektur von Tippfehlern (basierend auf einer zuvor erstellten Liste der gültigen Ausprägungen)
- Einheitliche Gross- bzw. Kleinschreibung (z.B. Müeller → MÜLLER)
- Umlaut-Normalisierung (z.B. MÜLLER → MUELLER)
- Segmentierung von zusammengesetzten Identifikatoren (auch *parsing* genannt), wie Doppelnamen, Adressen oder Geburtsdatum (so kann Adresse in einzelne Komponenten - Straße, Hausnummer, Postleitzahl und Stadt - extrahiert werden)
- Plausibilitätsüberprüfung (z.B. Länge und Existenz der Postleitzahl kontrollieren)

Im Laufe des Standardisierungsprozesses können auch einzelne fehlende Werte von vorhandenen Daten abgeleitet werden (z.B. fehlende Eingabe vom Geschlecht vom Vornamen (wenn eindeutig), oder Postleitzahl von der Stadt und Straße ableiten). In der Tabelle 2.2 sind die Datensets nach der Standardisierung abgebildet:

**Datenset A**

| Nachname | Vorname | Straße | PLZ | Stadt | Geschlecht | Geburtsdatum |
|---|---|---|---|---|---|---|
| Meier | Carsten | Salzachstr. 20 | 14163 | Berlin | m | 11-07-1964 |
| Hartman | Robert | Bismarckstr. 72 | 12157 | Berlin | m | 02-09-1957 |
| Wolff | Stefanie | Augsburger Weg 7 | 12309 | NA | w | 16-03-1999 |

**Datenset B**

| Name | Adresse | Geschlecht | GJahr | GMonat | GTstag |
|---|---|---|---|---|---|
| Stephanie Wolf | Augsburger W. 7, 13209 Berlin | w | 1999 | 03 | 16 |
| Hans Olaf Nay | Hofstrasse 18, 12157 BER | m | 1978 | 10 | 15 |
| Karsten Mayer | Salzachstr. 20 14163 Berlin | NA | 1964 | 11 | 07 |

**Tabelle 2.1:** Beispiel: zwei Datensets, die standardisiert werden sollen

**Datenset A**

| Name | VName1 | VNahme2 | Straße | Str.Typ | Haus Nr. | PLZ | Stadt | Geschl. | GJahr | GMonat | GTag |
|---|---|---|---|---|---|---|---|---|---|---|---|
| meier | carsten | | salzach | strasse | 20 | 14163 | Berlin | m | 1964 | 7 | 11 |
| hartman | robert | | bismarck | strasse | 72 | 12157 | Berlin | m | 1957 | 9 | 2 |
| wolff | stefanie | | augsburger | weg | 7 | 12309 | Berlin | w | 1999 | 3 | 16 |

**Datenset B**

| Name | VName1 | VNahme2 | Straße | Str.Typ | Haus Nr. | PLZ | Stadt | Geschl. | GJahr | GMonat | GTag |
|---|---|---|---|---|---|---|---|---|---|---|---|
| wolf | stephanie | | augsburger | weg | 7 | 12309 | Berlin | w | 1999 | 3 | 16 |
| nay | hans | olaf | hof | strasse | 18 | 12157 | Berlin | m | 1978 | 10 | 15 |
| mayer | karsten | | salzach | strasse | 20 | 14163 | Berlin | m | 1964 | 11 | 7 |

**Tabelle 2.2:** Datensets nach der Standardisierung

## 2.2 Blocking

Die bereinigten und standardisierten Daten sind nun bereit für das Linkage. Um zu überprüfen, ob ein Datensatz im Datenset A mit einem anderen Datensatz des Datenset B auf das gleiche Objekt verweist, müssen normalerweise jeder Datensatz von Datenset A mit jedem Datensatz des Datensets B verglichen werden. Das führt dazu, dass die Anzahl der Paarvergleiche quadratisch zur der Größe der Datensets ist. Für das Beispiel in der Tabelle 2.2 wären neun Vergleiche notwendig. Man kann sich leicht vorstellen, dass bei großen Datensätzen die Anzahl der benötigten Vergleiche schnell enorm groß werden kann, auch für einen maschinellen Vergleich (z.B. wenn beide Datensets aus ca. 1 Mio. Datensätzen bestehen, sind $10^{12}$ Paarvergleiche notwendig; nimmt man an, dass ca. 100000 Vergleiche pro Sekunde durchgeführt werden können, so würde man für den kompletten Vergleich von zwei solchen Datensets ca. 2778 Stunden oder 116 Tage brauchen (Christen [2012, S.27])). Der Großteil von diesen Paarvergleichen wird zwischen Datensätzen durchgeführt, die offensichtlich keine Matches sind. Um die große Anzahl an potentiellen Paarvergleichen zu reduzieren, werden die Datensets in kleine disjunkte Teilmengen, sogenannte Blocks, anhand von bestimmten Blocking-Regeln (oder *blocking key*) aufgeteilt. Nur die Datensätze von zwei Datensets, die in einem Block gelandet sind, werden miteinander verglichen. Als Beispiel für eine solche Blocking-Regel kann gleiche Postleitzahl oder gleicher phonetische Nachnamen dienen. Blocking-Regeln können auch mehrere Kriterien beinhalten.

## 2.3 Vergleich und Klassifikation

Im nächsten Schritt sollen die Datensätze innerhalb der Blocks miteinander verglichen werden. Dafür muss ein Maß eingeführt werden, das die Ähnlichkeit der von zwei Datensätzen widerspiegelt. Zur Berechnung von diesem Ähnlichkeitsmaßes gibt es zwei Ansätze - deterministischer und probabilistischer.

### 2.3.1 Deterministisches record linkage

Bei dem deterministischen Ansatz werden die Datensätze als Matches klassifiziert, wenn die Ausprägungen aller Linkage-Variablen exakt übereinstimmen. Man kann sich leicht vorstellen, dass die Anzahl solcher Fälle auch nach einer guten Datenaufbereitung ziemlich klein sein wird. Alternativ kann man für jedes Vergleichspaar die Summe $S$ berechnen, die Anzahl der exakten Übereinstimmungen über alle Merkmale wiedergibt:

$$S = \sum_{i=1}^{K} \quad \text{mit} \quad x_i = \begin{cases} 1, y_a^k = y_b^k \\ 0, y_a^k \neq y_b^k \end{cases} \tag{2.1}$$

wobei $y_a^k$ und $y_b^k$ Ausprägungen des Merkmales $y^k$ sind. Nachdem die Gewichte für alle Vergleichspaare berechnet wurden, soll ein Schwellenwert bestimmt werden, ab dem das Paar als Match klassifiziert wird. Oft werden zwei solche Schwellenwerte (oberer und unterer) betrachtet und die Fälle, die dazwischen liegen, werden manuell begutachtet und zugeordnet. Mögliche Verteilung von Matchgewichten ist in der Abbildung 2.2 dargestellt.

### 2.3.2 Probabilistisches record linkage

Die Basis-Idee des probabilistischen record linkage wurden von Newcombe et al. [1959] vorgeschlagen. Die Idee besteht darin, dass es bei zwei Datensätzen mit einem landess-

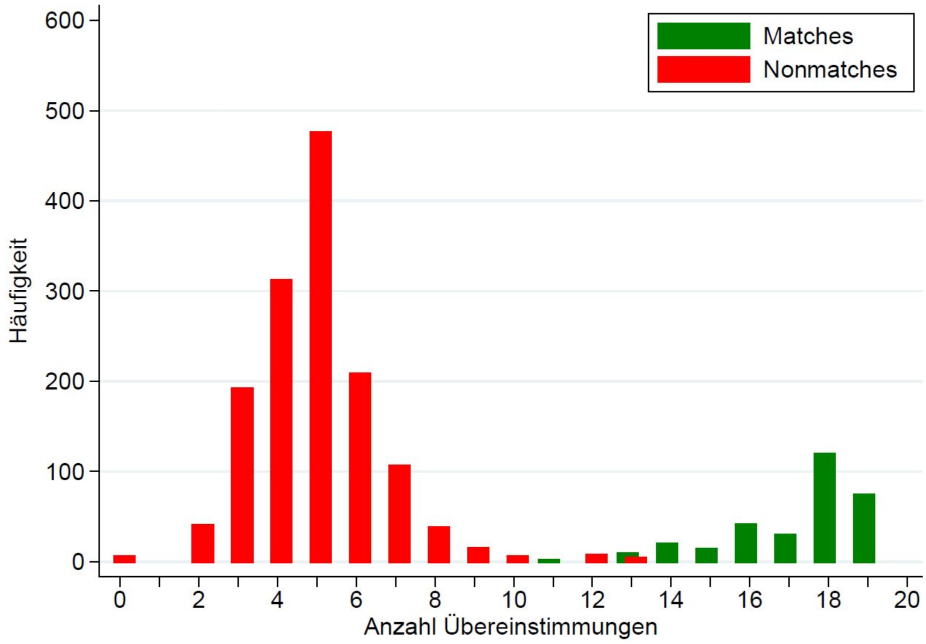


**Abbildung 2.2:** Verteilung Matchgewicht (Anzah Übereinstimmungen). Quelle: Gramlich [2014, S.14]

pezifisch seltenen Namen (z.B. Aiwasowski) wahrscheinlicher ist, dass diese Datensätze zusammengehören, als wenn es sich um einen häufigen Namen (z.B. Müller) handelt (Hentschel [2008, S.62-63]). Die Wahrscheinlichkeit steigt, wenn auch weitere Merkmale übereinstimmen. Diese Wahrscheinlichkeit kann man mit einem Übereinstimmungsgewicht bewerten, so wird der Übereinstimmung vom Nachnamen „Müller“kleineres Gewicht zugewiesen, als der Übereinstimmung von „Aiwasowski“. Fellegi and Sunter [1969] formalisierten diese Gedanken und entwickelten die Theorie von Record Linkage, die im Folgenden vorgestellt wird.

Seien $A$ und $B$ zwei Datensets, deren Elemente mit $a$ und $b$ bezeichnet werden. Wir nehmen an, dass manche Elemente gemeinsam für beide Datensets sind. Folglich kann man die Menge aller möglichen Paare

$$A \times B = \{(a, b); a \in A, b \in B\} \tag{2.2}$$

als die Vereinigung von zwei disjunkten Teilmengen

$$M = \{(a, b); a = b, a \in A, b \in B\} \tag{2.3}$$

und

$$U = \{(a,b); a \neq b, a \in A, b \in B\} \tag{2.4}$$

vorstellen, wobei sie als matched und unmatched bezeichnet werden. $M$ enthält somit alle Datensätze, die zum selben Objekt gehören, und $U$ alle anderen Paare. Aufgrund der Symmetrie gilt weiterhin, dass $(a,b) = (b,a)$, d.h. für zwei zusammengehörende Datensätze nur ein Match gibt. Man nimmt weiter an, dass es zwei zugrundeliegende datensatzgenerierende Prozesse für die Mengen $A$ und $B$ gibt. Das Resultat eines solchen Prozesses stellt einen Datensatz bzw. einen Merkmalsvektor für jedes Objekt dar. Dieser Merkmalsvektor enthält alle Werte (z.B. Name, Vorname, Adresse, Geburtsdatum usw.) zu einem bestimmten Objekt: $a = (a_1, ..., a_n)$ und $b = (b_1, ..., b_n)$. Der datengenerierende Prozess produziert auch manche Fehler oder fehlende Werte, deswegen können zwei nicht-gematchte Elemente von $A$ und $B$ identische Datensätze produzieren, und umgekehrt können zwei in Wirklichkeit identische Elemente unterschiedliche Datensätze aufweisen.

Wenn zwei Datensätze miteinander verglichen werden, wird für jedes Paar $(a,b)$ eine binärer Vergleichsvektor $\gamma$ generiert. In diesem Vektor können beliebige Ähnlichkeitskriterien kodiert werden, z.B. „Vorname stimmt überein“, „Geburtsdatum weicht maximal um 2 Jahre ab“, „Nachname ist ähnlich“. Also würde (1,1,0) bedeuten, dass die beiden ersten Kriterien übereinstimmen, das dritte jedoch nicht. Die Menge aller möglichen Ausprägungen von $\gamma$ bildet einen Vergleichsraum (*comparison space*) $\Gamma$. Man kann also sagen, dass eine Vergleichsfunktion $A \times B \to \Gamma$ Tupel $(a,b)$ auf einen Vektor $\gamma$ bildet.

Im letzten Schritt wird mit Hilfe der Entscheidungsfunktion (*linkage rule*) $\Gamma \to \{M, U, PM\}$ bestimmt, ob die betrachteten Datensätze in der Menge der Matches (M), Nicht-Matches(U) oder potenziellen Matches(PL) liegen. Als Grundlage für die Entscheidung dienen die Wahrscheinlichkeiten, dass $(a,b)$ in $M$ bzw. $U$ liegen. Formal lässt sich dass in Form von Agreement Ratio ausdrucken:

$$R = \frac{(\gamma \in \Gamma | (a,b) \in M)}{(\gamma \in \Gamma | (a,b) \in U)} \tag{2.5}$$

Somit ergibt sich folgende Entscheidungsfunktion:

$$\begin{cases} R \geq t_{obere} \Rightarrow (a, b) \rightarrow M \\ t_{untere} \leq R \leq t_{obere} \Rightarrow (a, b) \rightarrow PM \\ R \leq t_{untere} \Rightarrow (a, b) \rightarrow U \end{cases} \tag{2.6}$$

wobei $t_{obere}$ und $t_{untere}$ Schwellenwerte für Match und Nicht-Match darstellen.

Die Berechnung von bedingten Wahrscheinlichkeiten in Formel 2.5 ist eine der großen Herausforderungen von record linkage. Normallehrweise wird angenommen, dass diese Wahrscheinlichkeiten bedingt unabhängig für unterschiedliche Merkmale, die man beim Vergleich und Erstellung von $\gamma$ verwendet, sind. Unter dieser Annahme können individuelle Übereinstimmungsgewichte $w_i$ ($1 \leq i \leq K$, wobei $K$ die Anzahl an verwendete Linkage-Variablen ist) basieren auf m- und u-Wahrscheinlichkeiten berechnet werden:

$$m_i = P([a_i = b_i, a \in A, b \in B] | (a, b) \in M) \tag{2.7}$$

und

$$u_i = P([a_i = b_i, a \in A, b \in B] | (a, b) \in U) \tag{2.8}$$

wobei $a_i$ und $b_i$ die Werte der Variablen $i$ sind. Die Formel 2.7 gibt die Wahrscheinlichkeit an, dass die beiden Datensätze denselben Wert bei der Variablen $i$ aufweisen, wenn sie in Wirklichkeit zum selben Objekt gehören. $u_i$ dagegen ist die Wahrscheinlichkeit, dass die beiden Datensätze denselben Wert bei der Variablen $i$ aufweisen, obwohl sie zu unterschiedlichen Objekten gehören. Anhand von diesen zwei Parametern $m_i$ und $u_i$ wird das Gewicht für die Variable $i$ berechnet:

$$w_i = \begin{cases} \log_2(\frac{m_i}{u_i}) & \text{, falls} \quad a_i = b_i \\ \log_2(\frac{1-m_i}{1-u_i}) & \text{, falls} \quad a_i \neq b_i \end{cases} \tag{2.9}$$

Als kleines Rechenbeispiel betrachten wir die Variable „Geburtsmonat“. Man nimmt zunächst an, dass in 3% der Fälle diese Variable falsch erhoben wurde. Dies würde bedeuten, dass die Wahrscheinlichkeit, dass zwei Datensätze, die zu derselben Person gehören, gleiche Ausprägungen des Merkmals „Geburtsmonat“haben, $m_i = 0.97$ beträgt.

Die Wahrscheinlichkeit, dass zwei zusammengehörende Datensätze ungleichen Geburtsmonat haben, ist somit $1 - m_i = 1 - 0.97 = 0.03$. Die Wahrscheinlichkeit, dass zwei zufällig ausgewählte, nicht zusammengehörende Datensätze den gleichen Geburtsmonat aufweisen, beträgt $u_i = 1/12 = 0.083$. Die Wahrscheinlichkeit, dass zwei zufällig ausgewählte, nicht zusammengehörige Datensätze unterschiedliche Geburtsmonate aufweisen, ist $1 - u_i = 1 - 0.083 = 0.917$. So kann man dazugehörige Gewichte berechnen:

$$w_i = \begin{cases} \log_2(\frac{0.97}{0.083}) = 3.54 & \text{, falls} \quad a_i = b_i \\ \log_2(\frac{0.03}{0.917}) = -4.92 & \text{, falls} \quad a_i \neq b_i \end{cases}$$

Da die bedingte Unabhängigkeit angenommen wurde, bildet sich das Gesamtgewicht für jedes Vergleichspaar als die Summe von einzelnen Gewichten über alle Linkage-Variablen:

$$w = \sum_{i=1}^{K} w_i \tag{2.10}$$

Die Verteilung der Übereinstimmungsgewichten nimmt normalerweise eine unsymmetrische U-förmige Gestalt auf.

Im vorletzten Schritt der Datenanalyse erfolgt eine manuelle Zuordnung der potentiellen Matches zu den Gruppen der Matches bzw. Nicht-Matches. Danach werden die gematchten Datensätze in einem Master-Datenset aggregiert.

# Literaturverzeichnis


Peter Christen. *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection.* Springer Science & Business Media, 2012.

Halbert L Dunn. Record linkage*. *American Journal of Public Health and the Nations Health*, 36(12):1412–1416, 1946.

Ivan P Fellegi and Alan B Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, 1969.

Tobias Gramlich. strokes record linkage der schlaganfälle in hessen 2007-2010. Technical report, 2014.

Stefan Hentschel. *Das manual der epidemiologischen Krebsregistrierung.* Zuckschwerdt, 2008.

Daniel Nasseh. *Einsatz und Optimierung einer überwachten Klassifizierungsmethode im Kontext eines Privacy-Preserving-Record-Linkage.* PhD thesis, lmu, 2014.

Howard B Newcombe, James M Kennedy, SJ Axford, and Allison P James. Automatic linkage of vital records computers can be used to extract"follow-upßtatistics of families from files of routine records. *Science*, 130(3381):954–959, 1959.