Bayerisches Landesamt für Statistik und Datenverarbeitung

Mit Bayern rechnen.

RECORD-LINKAGE IM zensus 2011

Der maschinelle Namensabgleich der Haushaltegenerierung

Marco Reisch
Bayerisches Landesamt für Statistik und Datenverarbeitung

1. Workshop des German RLC **Nürnberg, 04.-05.04.2012**

Gliederung

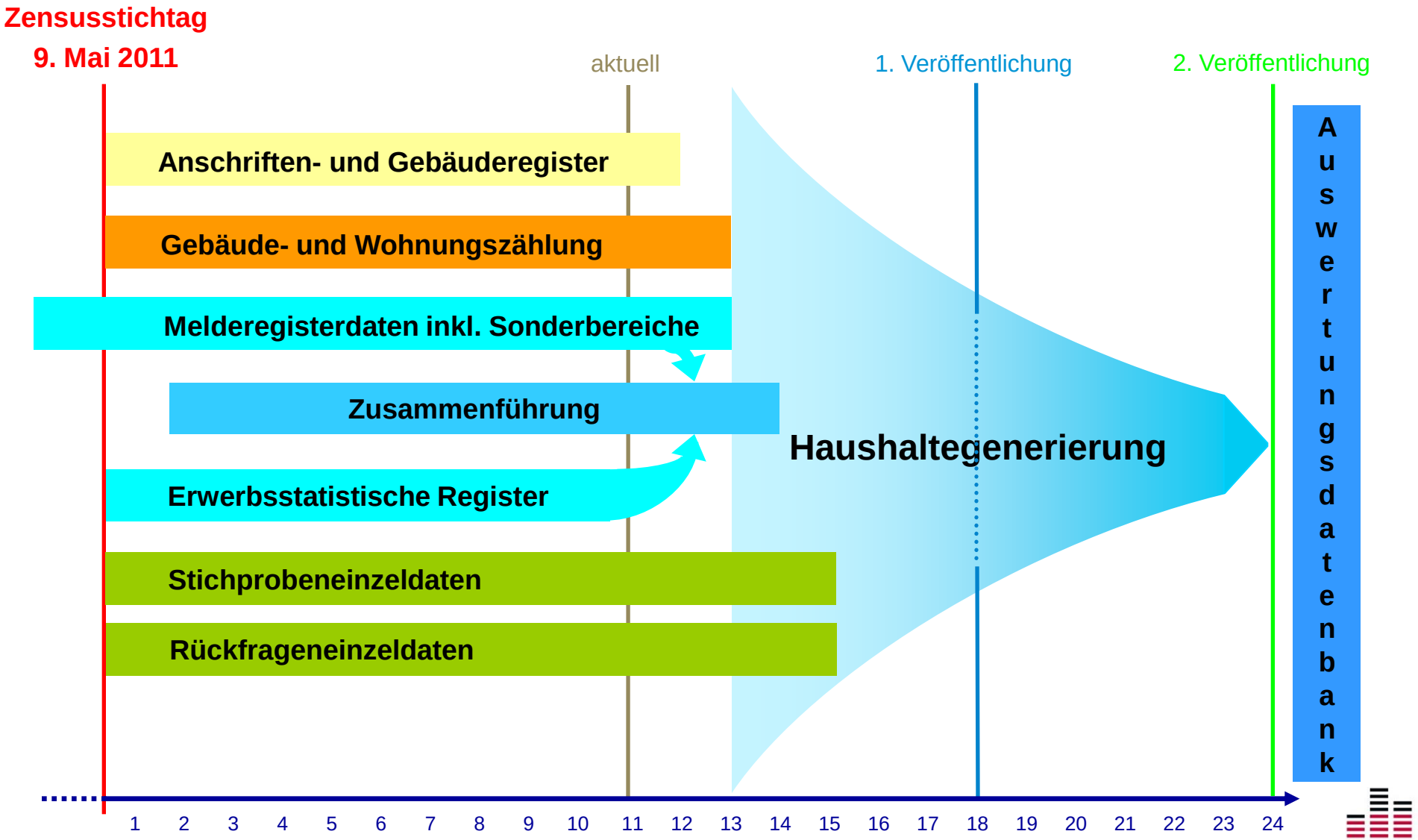
- **Datenbasis und Zeitmodell des Zensus 2011**
- **Die Haushaltegenerierung**
- **Der maschinelle Namensabgleich**

Datenbasis des Zensus 2011

Registergestützter Zensus auf Basis unterschiedlicher Datenquellen

- **Basis aller Erhebungsteile bildet das Anschriften- und Gebäuderegister**
  alle Anschriften Deutschlands mit Wohnraum, gewonnen aus unterschiedlichen Quellen
- **Registerdaten aus den Einwohnermeldeämtern**
  - **erweitert um erwerbsstatistische Daten der BA und der öffentlichen Arbeitgeber**
  - **bereinigt um die Ergebnisse der Befragung in Wohnheimen und Gemeinschaftsunterkünften gemäß § 8 ZensG 2011 und der Befragung zur Klärung des Wohnsitzes gemäß § 15 ZensG 2011**
- **Daten aus der Gebäude- und Wohnungszählung**
- **Ergänzende Befragungen zur Feststellung der amtlichen Einwohnerzahl und der Ermittlung von Zusatzinformationen**
  - **Haushaltebefragung gemäß § 7 ZensG 2011**
  - **Klärung von Unstimmigkeiten gemäß § 16 ZensG 2011**

Das Zeitmodell des Zensus 2011



Gliederung

- **Datenbasis des Zensus 2011**
- **Die Haushaltegenerierung**
- **Der maschinelle Namensabgleich**

Die Haushaltegenerierung - Ziele

Registergestützter Zensus auf Basis unterschiedlicher Datenquellen

→ **Statistisches Verfahren benötigt, welches…**
(Zielsetzungen der Haushaltegenerierung)

- einen zensustypischen Datensatz zur Auswertungen von übergreifenden Merkmalskombinationen aus den verschiedenen Erhebungsteilen auf fachlich und regional tief gegliederter Ebene erlaubt
- den Registerpersonenbestand mit Hilfe der Erkenntnisse aus der Haushaltebefragung korrigiert um Einzeldaten für flexible Auswertungen bereitzustellen
- Daten zur Zahl und Struktur von Haushalten (Wohnhaushalte) ermittelt

Die Haushaltegenerierung - Das Modell

Gliederung

- **Datenbasis des Zensus 2011**
- **Die Haushaltegenerierung**
- **Der maschinelle Namensabgleich**

Konzept der Zusammenführung

Verfahrensschritte des Datenabgleichs

- **Bestimmung zu identifizierender Merkmale**
- **Vergleich der Datensatzpaare**
- **Klassifikation der Datensatzpaare in identisch/nicht identisch**

  Komponenten einer festen Bewertungsregel für die Merkmale m_j

  - **Vergleichsfunktionen $f_j(m_j)$**
  - **Bewertungsfunktion $\lambda(f_j(m_j))$**
  - **Entscheidungsfunktion $\delta(\lambda)$**

Konzept der Zusammenführung

Bestimmung zu identifizierender Merkmale

Art der Wohnungsnutzung aus der GWZ

Merkmale der Wohnungsnutzer aus der GWZ jeweils für 1. und 2. Wohnungsnutzer	Merkmale der gemeldeten Personen im Melderegister (Personensatz)
Nachname (einschl. Namensbestandteilen und Doktortitel)	Familienname einschl. Namensbestandteilen und Doktortitel
	Geburtsname einschl. Namensbestandteilen
	Ehename einschl. Namensbestandteilen
	Familienname vor Änderung einschl. Namensbestandteilen
	Familienname des Partners/Vorfahren/Nachkommen
Vornamen	Vornamen
	Rufnamen
	Vornamen des Partners/Vorfahren/Nachkommen

Konzept der Zusammenführung

Vergleich der Datensatzpaare: vorbereitende Maßnahmen

- **Ersetzen von Sonderzeichen durch Blanks**
  Ausnahme: Wildcard-Symbol der Beleglesung ‚*‘
- **Ersetzen von überzähligen Blanks**
  Beispiel: doppelte Blanks, Blanks am Anfang oder Ende von Namen
- **Umsetzen von Umlauten**
  Beispiel: ä → ae
- **Umsetzen von Zeichen/Zeichenkombinationen**
  Beispiel: ß → ss, ph → f, th → t
- **Entfernen von Akzenten**
  Beispiel: é → e, ç → c
- **Umsetzen aller Klein- in Großbuchstaben**

Konzept der Zusammenführung

Vergleich der Datensatzpaare: Problemstellung

- **Problem: Namen häufig nicht vollständig identisch**
  - **Fehlerhafte Angaben in der GWZ**
  - **Fehler bei der Datenerfassung (Beleglesung)**
    Falsch bzw. nicht erkannte Zeichen
  - **Fehler in den Melderegistern**
- **Ziel: Algorithmus, der auch „Ähnlichkeiten“ von Namen aufdeckt**
- **Anforderungen an den Algorithmus**
  - **Maximierung der Trefferquote**
    (= Minimierung des manuellen Aufwands)
  - **Minimierung von Falschzuordnungen**

Konzept der Zusammenführung

Vergleich der Datensatzpaare: Das Phonetikmodul

- Vergleich von Zeichenketten positionsgenau

  GWZ: BANDISCH

  MR: BAUDISCH

- Vergleich innerhalb eines Suchraums

  GWZ: BIERMANN

  MR: BIRMANN

  - Suchraum begrenzt

Konzept der Zusammenführung

Vergleich der Datensatzpaare: Das Phonetikmodul

- **Vergleich innerhalb eines Suchraums**

  GWZ: BIERMANN

  MR: BIRMANN

- **Bewertung des Abgleichs**

$$\text{Bewertungsquotient} = \frac{\text{Summe der Laengen der gemeinsamen Teilstrings}}{\text{Laenge des laengeren Namens}}$$

$$\text{Bewertungsquotient} = \frac{2+5}{8} = 0{,}875$$

Konzept der Zusammenführung

Vergleich der Datensatzpaare: Das Phonetikmodul

- **Beispiele:**

Gesuchter String	Vergleichsstring	Länge der gemeinsamen Teilstrings	Länge des längeren Namens	Bewertungsquotient
Franziska	Francisca	4+2	9	0,67
Marie	Maria	4	5	0,8
Jeanette	Jeannette	4+4	9	0,89
Gretchen	Grete	4	8	0,5
Erwin	Ervin	2+2	5	0,8
Drakomena	Draomina	3+2+2	9	0,78
Rieki	Rilki	2+2	5	0,8
Atanassioni	Atanasiou	6+2	11	0,73
Bandisch	Baudisch	2+5	8	0,88
Lieskovsky	Lieszkovszky	4+4+2	12	0,83

Konzept der Zusammenführung

Klassifikation der Datensatzpaare

- **Vergleich zweier Datenbestände:**

  paarweiser Vergleich zwischen jedem Datensatz des einen Datenbestandes mit jedem Datensatz des anderen Datenbestandes

  → wenig praktikabel

- **Prinzip der sukzessiven Massenreduktion:**
  - **ersten Satz des einen Datenbestandes mit jedem Datensatz des anderen Datenbestandes vergleichen**
  - **Klassifikation in identisch/nicht identisch**
  - **Fortfahren mit nächstem Satz des einen Datenbestandes mit den verbleibenden Sätzen des anderen Datenbestandes**

Konzept der Zusammenführung

Klassifikation der Datensatzpaare

- **Methodisches Problem des Prinzips der sukzessiven Massenreduktion**

GWZ		
Lfd. Nr.	Familienname	Vorname
1	Müller	Petra
2	Müller	Peter

Melderegister		
Lfd. Nr.	Familienname	Vorname
1	Miller	Hans-Peter
2	Müller	Peter

- **Lösung: Verwendung mehrerer hierarchisch abgestufter Bewertungsregeln**

Konzept der Zusammenführung

Klassifikation der Datensatzpaare: Hierarchisches Stufenmodell

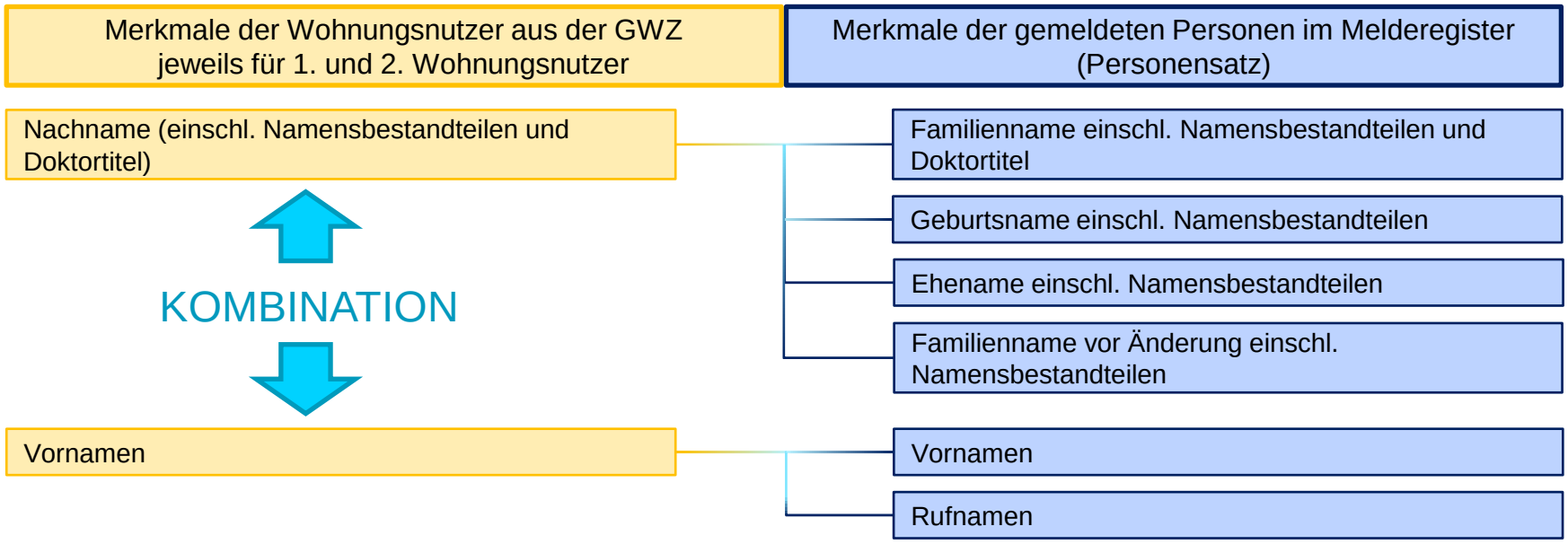
- Zunächst Vergleich zweier Datenbestände mit sehr restriktiver Bewertungsregel nach dem Prinzip der sukzessiven Massenreduktion
- Im weiteren werden die Bewertungsregeln zunehmend „weicher“
- Fortfahren mit nächstem Satz des einen Datenbestandes mit den verbleibenden Sätzen des anderen Datenbestandes
  - Bei voller Übereinstimmung zweier Namen Bewertung von 1
  - Bei Namensähnlichkeiten Bewertung mittels Phonetikmodul, Ähnlichkeitsmaß im Intervall [0, 1]

  Schwellwert für Klassifikation in identisch situationsbedingt (einfache/doppelte Phonetik, Ähnlichkeitsbewertung von Vornamen/Nachnamen)

Konzept der Zusammenführung

Klassifikation der Datensatzpaare: Hierarchisches Stufenmodell

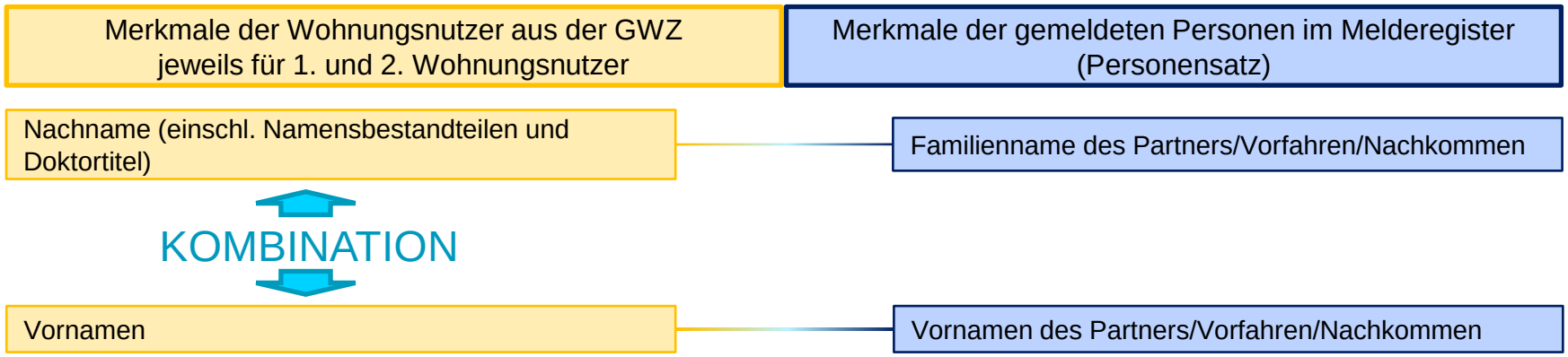
- **Gesamtablauf gliedert sich in drei Stufen**
  - **1. Stufe: Namensabgleich im engeren Sinne (46 Unterstufen)**



Konzept der Zusammenführung

Klassifikation der Datensatzpaare: Hierarchisches Stufenmodell

- **Gesamtablauf gliedert sich in drei Stufen**
  - **2. Stufe: Namensabgleich über Verzeigerungen (11 Unterstufen)**



Konzept der Zusammenführung

Klassifikation der Datensatzpaare: Hierarchisches Stufenmodell

- **Gesamtablauf gliedert sich in drei Stufen**
  - **3. Stufe: Unechter Namensabgleich (4 Unterstufen)**

Merkmale der Wohnungsnutzer aus der GWZ jeweils für 1. und 2. Wohnungsnutzer	Merkmale der gemeldeten Personen im Melderegister (Personensatz)
Nachname (einschl. Namensbestandteilen und Doktortitel) *Voraussetzung: darf nur einmal an der Anschrift vorhanden sein*	Familienname einschl. Namensbestandteilen und Doktortitel

Konzept der Zusammenführung

Klassifikation der Datensatzpaare: Hierarchisches Stufenmodell

- **Unterstufen der Stufen 1 und 2 zum Teil verstärkt durch Synonym-Suche**
  - **Teilweise starke Abweichungen zwischen den Melderegisterangaben zum Vornamen und den GWZ-Angaben (Kurzformen)**
  - **Synonymliste für Vornamen hinterlegt z.B. Josef ↔ Sepp, Ulrike ↔ Uschi**
  - **Ausschließlich Unterstufen ohne Ähnlichkeits- oder Teilvornamen-Suche**

  - **betrifft 11 Unterstufen der Stufe 1 → 57 ‚Unterstufen'**
  - **betrifft 2 Unterstufen der Stufe 2 → 13 ‚Unterstufen'**

Konzept der Zusammenführung

Klassifikation der Datensatzpaare: Hierarchisches Stufenmodell

- **Unterstufen der Stufe 1 zum Teil verstärkt durch Umlaut-Suche**
  - **Beleglesung erkennt in den meisten Fällen Umlaute nicht korrekt → a statt ä, o statt ö und u statt ü**
  - **Bereinigung der standardisierten Namen auf beiden Seiten um Umlaute → AE → A, OE → O und UE → U**
  - **Keine Kombination mit Synonym-Suche**

- **betrifft alle 46 Unterstufen der Stufe 1 → 103 ‚Unterstufen'**

Konzept der Zusammenführung - Testergebnisse

- **Zensustest**
  - Auswertung der maschinellen Verknüpfung von 1683 paarigen Fällen
  - Richtigzuordnungen bei 99,3 % (Rest keine Zuordnung)
- **Simulation eines Beleglese-Szenarios**
  - Beleglesung von rund 18900 handschriftliche Wohnungsnutzerangaben
  - Betrachtung auch nichtpaariger Fälle
  - Richtigzuordnungen bei 88 %, Falschzuordnungen bei 0,0019%
- **Testrechnungen mit unplausibilisierten Rohdaten**
  - rund 49,5 Mio. Wohnungsnutzerangaben bei etwa 40,5 Mio. Wohnungsdatensätzen
  - Erhebungsformen: Beleg, IDEV, Core-Meldungen
  - Trefferquote bei 81,6 %

Abschließendes Beispiel: Doppelte Phonetik

Hinreichende Ähnlichkeit der Familien- und Vornamen (Stufe 1)

GWZ		
Lfd. Nr.	Familienname	Vorname
1	ZIMMERMANN	JOHANNES

Melderegister		
Lfd. Nr.	Familienname	Vorname
1	ZIMMERER	HANS
2	SEMMERMANN	JOHANN
3	ZIMMER	HANNES

- **Schwellwerte für Abgleich mit doppelter Phonetik**
  - **Schwellwert Familiennamenbewertung: 0,6**
  - **Schwellwert Vornamenbewertung: 0,5**
  - **Schwellwert Gesamtbewertung: 0,6**

Abschließendes Beispiel: Doppelte Phonetik

Hinreichende Ähnlichkeit der Familien- und Vornamen (Stufe 1)

- **Vergleichsfunktionen:**
  - **Eine mögliche Zusammenführung soll nur bei Datensatzpaaren erfolgen, bei denen die phonetische Übereinstimmung beider Merkmale mindestens die geforderten Schwellwerte beträgt.**
  - **Vergleichsfunktion Familiennamenbewertung**

$$f_1(F) = \begin{cases} p(F), & \text{falls } p(F) \geq 0{,}6 \\ 0, & \text{sonst} \end{cases}$$

  - **Vergleichsfunktion Vornamenbewertung**

$$f_2(V) = \begin{cases} p(V), & \text{falls } p(V) \geq 0{,}5 \\ 0, & \text{sonst} \end{cases}$$

Abschließendes Beispiel: Doppelte Phonetik

Hinreichende Ähnlichkeit der Familien- und Vornamen (Stufe 1)

- **Bewertungsfunktion:**
  - **Datensatzpaare erhalten eine Gesamtbewertung > 0, wenn das arithmetische Mittel der einzelnen Namensähnlichkeiten mindestens dem geforderten Schwellwert für die Gesamtbewertung entspricht.**

$$\lambda(f_1(F), f_2(V)) = \begin{cases} 0{,}5(f_1(F) + f_2(V)), & \text{falls } 0{,}5(f_1(F) + f_2(V)) \geq 0{,}6 \\ & \text{und } f_1(F), f_2(V) > 0 \\ 0, & \text{sonst} \end{cases}$$

Abschließendes Beispiel: Doppelte Phonetik

Hinreichende Ähnlichkeit der Familien- und Vornamen (Stufe 1)

- **Entscheidungsfunktion:**
  - **Das Datensatzpaar mit der höchsten Gesamtbewertung der Übereinstimmung (> 0) wird als identisch klassifiziert.**

$$\delta(\lambda) = \begin{cases} \text{identisch}, & \text{falls } \lambda > 0 \text{ und } \lambda = \max_i \lambda_i \\ \text{Nicht identisch}, & \text{sonst} \end{cases}$$

Abschließendes Beispiel: Doppelte Phonetik

Hinreichende Ähnlichkeit der Familien- und Vornamen (Stufe 1)

- **Phonetikmodul liefert folgende Ähnlichkeitsmessungen:**

Paar	GWZ		Melderegister		phon. Ähnlichkeit	
	Familienname	Vorname	Familienname	Vorname	p(F)	p(V)
(a, b_1)	ZIMMERMANN	JOHANNES	ZIMMERER	HANS	0,6	0,375
(a, b_2)	ZIMMERMANN	JOHANNES	SEMMERMANN	JOHANN	0,8	0,75
(a, b_3)	ZIMMERMANN	JOHANNES	ZIMMER	HANNES	0,6	0,75

- **Vergleich, Bewertung und Entscheidung:**

Paar	Vergleich		Bewertung	Bewertung
	p(F)	p(V)	λ(f1 (F), f2 (V))	$\delta(\lambda)$
(a, b_1)	0,6	0	0	nicht identisch
(a, b_2)	0,8	0,75	0,775	identisch
(a, b_3)	0,6	0,75	0,675	nicht identisch

→ **Johannes Zimmermann wird mit Johann Semmermann zusammengeführt.**

Vielen Dank
für Ihre Aufmerksamkeit!

Marco Reisch
Tel.: 089 / 2119 500 244
E-Mail: Marco.Reisch@lfstad.bayern.de

zensus 2011