teren kräftigen Zuwachs von 6,9% an eine inzwischen acht Jahre anhaltende Folge hoher Steigerungsraten anschließen. Seit dem Rezessionsjahr 1993 betrug der Leistungsanstieg hier immerhin bereits 55,3%. Ein noch etwas stärkeres Wachstum verzeichnete lediglich das Kreditgewerbe, welches hier zusammengefaßt mit dem Versicherungsgewerbe nachgewiesen wird. Zusammen konnte das Kredit- und Versicherungsgewerbe – weitgehend unbeschadet der seinerzeitigen Rezession – seine Wertschöpfung innerhalb der letzten zehn Jahre um 77,5% steigern. Auch im Vorjahresvergleich wurde hier 2001 mit 5,6% wieder der mit Abstand höchste Zuwachs unter allen Hauptbereichen der Wirtschaft erzielt.

Natürlich hatten auch die übrigen Dienstleistungsbereiche wieder eine Steigerung ihrer Wirtschaftsleistung aufzuweisen. So leisteten die sogenannten Unternehmensdienstleister (einschließlich Grundstückswesen und Vermietung) mit 2,2% erneut einen beträchtlichen Beitrag zum gesamtwirtschaftlichen Wachstum. Im Vorjahr war der Zuwachs allerdings noch doppelt so hoch ausgefallen. Mit einem leichten Zuwachs konnten sich demgegenüber die öffentlichen und privaten Dienstleister behaupten. Allerdings halbierte sich auch hier die reale Zuwachsrate von 2,4 auf nun 1,2%.

Auf mittelfristige und längere Sicht kommt dem Tertiären Sektor gesamtwirtschaftlich eine stabilisierende und auf dauerhafte Expansion gerichtete Funktion zu, die rein oberflächlich betrachtet in einem gewissen Kontrast zu den teilweise drastischen Konjunkturschwankungen im Sekundären Sektor steht. Diese Entwicklung kann allerdings auch als Zeichen einer weiter wachsenden sektoralen, regionalen, aber auch globalen Verflechtung verstanden werden, welche die arbeitsteilige Form des Wirtschaftens nachhaltig forciert. Als Fakt bleibt dennoch, daß der Sekundäre Sektor als Kernbereich des Wirtschaftsgeschehens auch den „Nährboden“ für sehr viele, insbesondere „industrienahe“, Dienstleistungen darstellt, so daß dem Verarbeitenden Gewerbe sogar in doppelter Hinsicht Bedeutung im Wirtschaftsgeschehen zukommt: erstens durch den eigenen, wenngleich schwankenden und langfristig moderaten Wachstumsbeitrag und zweitens durch zusätzliche Mobilisierung von unternehmensbezogenen Dienstleistungen.

Dr. Franz Kohlhuber

---

1) Angaben für die beiden deutschen Teilregionen jeweils ohne Berlin.

2) Der Arbeitskreis „Volkswirtschaftliche Gesamtrechnungen der Länder“, dem alle Statistischen Landesämter angehören, führte bereits im Februar eine erste Berechnung des Wirtschaftswachstums in den deutschen Bundesländern durch, welche allerdings noch auf einer Auswertung statistischer Daten des Zeitraumes Januar bis Oktober basierte und somit als Hochrechnung für das gesamte Jahr zu verstehen war. Der inzwischen vorgelegten zweiten Berechnung liegen nunmehr vollständige Jahresdaten sowie auch sektoral eine breitere statistische Datenbasis zugrunde.

3) Die Zahlen für die neuen Länder sind aufgrund der Sonderentwicklung im Anschluß an die Wiedervereinigung hier nur bedingt vergleichbar.

4) Berlin: – 6000 Beschäftigte bzw. – 0,4%.

5) Der seit vielen Jahren bestehende Trend zu reduzierten Arbeitszeiten bzw. zur Teilzeitbeschäftigung im hergebrachten Sinn ist in dieser Abschätzung noch nicht berücksichtigt. Er würde die Diskrepanz noch etwas vergrößern, gemessen am Effekt der geringfügigen Beschäftigungsverhältnisse allerdings nur marginal.

6) Gemessen anhand des Preisindex für die Gesamtlebenshaltung aller privaten Haushalte in Deutschland.

Zusammenführung von Datenbeständen ohne numerische Identifikatoren

– ein Verfahren im Rahmen der Testuntersuchungen zu einem registergestützten Zensus –

Zukünftig sollen in der amtlichen Statistik die benötigten Daten anstelle der herkömmlichen Befragungen verstärkt auch aus Verwaltungsregistern gewonnen werden. Da diese in der Regel die statistischen Informationen nur teilweise enthalten, ergibt sich die Notwendigkeit, Daten aus verschiedenen Datenbeständen maschinell zu einem statistisch auswertbaren Gesamtbestand zusammenzuführen. – Im vorliegenden Beitrag wird zunächst auf die methodischen Grundlagen der Zusammenführung von Datenbeständen eingegangen. Vor allem Zusammenführungen ohne Identifikatoren werfen erhebliche methodische Schwierigkeiten auf, da hier der Zusammenhang unterschiedlich strukturierter Datensätze anhand von Ähnlichkeiten gemeinsamer Merkmale überprüft werden muß. Diese Prüfungen beinhalten neben der Messung von Ähnlichkeiten der Merkmalsausprägungen – insbesondere von Namen – auch das Aufstellen von Bewertungs- und Entscheidungsregeln, wann Datensätze als zusammengehörig betrachtet werden sollen und wann nicht. – Derzeit wird in der amtlichen Statistik anhand umfangreicher Testuntersuchungen geprüft, ob eine künftige Volkszählung auf der Basis von Verwaltungsregistern durchgeführt werden kann. Im Rahmen dieses Zensustests sollen unter anderem die Daten aus dem Melderegister und einer postalisch durchgeführten Gebäude- und Wohnungszählung zusammengeführt werden. Für diese spezielle Anwendung wurde im Bayerischen Landesamt für Statistik und Datenverarbeitung ein Verfahren konzipiert, das im wesentlichen auf einer neu entwickelten Methode des Zeichenkettenvergleichs zur Quantifizierung von Namensähnlichkeiten sowie einem hierarchisch strukturierten Stufenmodell zur Bewertung von Datensatzübereinstimmungen basiert. – Für das entwickelte Verfahren liegen inzwischen erste vielversprechende Testergebnisse vor.

1. Einführung

Registernutzung in der amtlichen Statistik

Die amtliche Statistik in Deutschland steht an einem Wendepunkt. Während – vor allem in den Wirtschafts- und Sozialstatistiken – die Gewinnung der Informationen bisher weitgehend durch die direkte Befragung von Bürgern oder Unternehmen (Primärstatistik) erfolgte, sollen künftig zunehmend Verwaltungsregister für statistische Auswertungen herangezogen werden (Registerstatistik).

Die Vorteile dieser Substitution von Befragung durch Auswertung vorhandener Daten liegen auf der Hand:

- Entlastung von Bürgern und Unternehmen von statistischer Auskunftserteilung,
- höhere Akzeptanz der amtlichen Statistik,
- Kosteneinsparungen in den statistischen Ämtern.

Bereits seit einer Reihe von Jahren unternimmt die amtliche Statistik Anstrengungen in dieser Richtung. So ist es in Bayern erfolgreich gelungen, die Belastungen landwirtschaftlicher Betriebe mit Statistikaufgaben durch Nutzung von Datenbeständen der Landwirtschaftsverwaltung zu reduzieren. Mit dem derzeit im Aufbau befindlichen Unternehmensregister, in das auch Daten der Finanzverwaltungen, der Bundesanstalt für Arbeit und der Kammern einfließen, soll mittelfristig eine merkliche Verringerung primärstatistischer Erhebungen bei Betrieben und Unternehmen aller Branchen erzielt werden. Mit den Testuntersuchungen für einen registergestützten Zensus wird derzeit geprüft, ob es auch bei der wichtigsten Aufgabe der amtlichen Statistik, der Volkszählung, künftig möglich sein wird, auf eine Befragung der Bürger ganz oder teilweise zu verzichten.

Die Umstellung einer Primärstatistik auf ein registergestütztes Verfahren stellt die amtliche Statistik vor eine Vielzahl von organisatorischen, konzeptionellen und juristischen Schwierigkeiten[1)]. Ein Problembereich beim Umstieg auf die Registerstatistik, mit dem sich der vorliegende Beitrag beschäftigt, ist die Zusammenführung von Datenbeständen aus verschiedenen Dateien, in der Literatur auch als Data-Merging, Data-Matching oder Data-Integration bezeichnet. Bevor im folgenden ein vom Bayerischen Landesamt für Statistik und Datenverarbeitung hierfür im Rahmen des Zensustests entwickeltes Verfahren vorgestellt wird, soll zunächst auf die Notwendigkeit von Datenzusammenführungen bei der Registerstatistik sowie die theoretischen Grundlagen eingegangen werden.

Registerstatistik erfordert die Zusammenführung von Datenbeständen

Es ist eher die Ausnahme, daß alle für die Durchführung einer amtlichen Statistik erforderlichen Daten aus einem einzigen Verwaltungsregister übernommen werden können, wie dies beispielsweise bei der Statistik der Bewilligungen im sozialen Wohnungsbau der Fall ist. Bei der Verwendung von Verwaltungsdaten für statistische Zwecke sieht man sich derzeit in Deutschland vielmehr mit der Situation konfrontiert, daß die einzelnen Verwaltungsregister aufgrund ihrer ursprünglichen Zweckbestimmung nur einen Teil der für die Statistik benötigten Daten beinhalten. Um diese Teilinformationen dennoch nutzen zu können, ist es notwendig, die Daten eines Registers ggf. mit den Daten anderer Register und/oder – sofern in Registern nicht vorhandene Angaben benötigt werden – mit den Daten einer ergänzenden Primärstatistik zu kombinieren.

Betrachten wir am fiktiven Beispiel einer Wohnungsstatistik zunächst den Fall, daß die potentiellen Datenquellen (Register) jeweils nur einen Teil der für eine Statistik benötigten Informationen (Merkmale) aufweisen. Es stehen zwei Register A und B zur Verfügung; A enthält Personen- und Haushaltsdaten, B enthält Daten zu Wohnungen. Wertet man die beiden Datenbestände jeweils für sich aus, so enthält man zwar statistische Informationen zur Zahl und Struktur der Haushalte und der Wohnungen, Informationen darüber, wie die Haushalte mit Wohnraum versorgt sind, lassen sich hieraus jedoch nicht ermitteln. Diese lassen sich erst durch eine Zusammenführung (Verknüpfung) der Einzeldaten beider Datenquellen, also der Zuordnung der einzelnen Haushalte in die jeweils von ihnen bewohnten Wohnungen, gewinnen.

Anders stellt sich die Situation dar, wenn die verwendeten Register nur Daten über eine Teilmasse der zu erhebenden statistischen Einheiten enthalten. Zusammenführung bedeutet hier zunächst die Vereinigung der Register – und ggf. primärstatistisch gewonnener Daten – zu einem Gesamtdatenbestand, der dann entsprechend statistisch ausgewertet werden kann. Dabei ist allerdings zu berücksichtigen, daß die einzelnen Teilmassen häufig nicht disjunkt sind, also einzelne Erhebungseinheiten in mehreren Datenbeständen vorkommen können. So kann bspw. eine Person gleichzeitig sozialversicherungspflichtig beschäftigt und arbeitslos sein und folglich sowohl in der Datei der sozialversicherungspflichtig Beschäftigten als auch in der Arbeitslosendatei der Bundesanstalt für Arbeit registriert sein. Ohne eine entsprechende Bereinigung dieser Dubletten würden solche Personen bei einer Zusammenführung der Datenbestände doppelt gezählt werden.

Ob Verknüpfung von Merkmalswerten oder Verschmelzung von Teilmassen, im Kern beinhaltet die Zusammenführung immer die Prüfung von Datensatzpaaren unterschiedlicher Datenbestände auf ihre Zusammengehörigkeit (im folgenden als Identität bezeichnet). Ein solcher Datenabgleich setzt voraus, daß in den zu vergleichenden Datensätzen gemeinsame Merkmale enthalten sind, anhand derer die Identität zwischen den Datensatzpaaren erkannt werden kann. Hinsichtlich dieser Voraussetzung lassen sich Zusammenführungen mit und ohne Identifikatoren unterscheiden.

2. Methodische Grundlagen der Zusammenführung von Datenbeständen

Zusammenführung von Datenbeständen mit Identifikatoren

Unter einem Identifikator (sog. Primärschlüssel) versteht man ein in den zusammenzuführenden Dateien enthaltenes eindeutiges, einheitliches, i.d.R. numerisches Kennzeichen (Hilfsmerkmal) der Einheiten. Da gleiche Identifikatoren gleiche Einheiten bezeichnen, läßt sich die Identität zweier Einheiten zweifelsfrei (deterministisch) bestimmen.

Numerische Identifikatoren werden in der amtlichen Statistik zur Zusammenführung von Ergebnissen aus unterschiedlichen Erhebungsteilen verwendet. Hierbei werden bereits im Vorfeld an die verschiedenen Erhebungsteile eindeutige Identifikationsnummern (i.d.R. durch Aufdruck auf die Erhebungsbogen) vergeben, über die dann die Ergebnisse der Teilerhebungen zu einem Gesamtbestand für die Ergebniserstellung zusammengeführt werden. Ein Beispiel hierfür ist die Bautätigkeitsstatistik, bei der die Angaben zur Genehmigung eines Bauvorhabens (Baugenehmigungsstatistik) zu einem späteren Zeitpunkt mit den Angaben zur Fertigstellung des Bauwerks (Baufertigstellungsstatistik) über eine Identifikationsnummer zu einem Datensatz verknüpft werden.

Für eine Nutzung von Verwaltungsregistern stehen derzeit in Deutschland keine numerischen Identifikatoren zur Verfügung. So hat das Bundesverfassungsgericht in

seinem Volkszählungsurteil[2)] 1983 für Personendaten die Einführung eines einheitlichen, für alle Register und Dateien geltenden Personenkennzeichens oder dessen Substituts untersagt. Das Bundesverfassungsgericht sieht die Einführung von Personenkennziffern als „... den entscheidenden Schritt, den einzelnen Bürger in seiner ganzen Persönlichkeit zu registrieren und zu katalogisieren." In den skandinavischen Ländern werden diese Gefahren offenbar nicht als derart gravierend eingeschätzt. Hier wurden bereits vor mehr als zehn Jahren Personenkennziffern eingeführt. Damit wurde eine wesentliche Voraussetzung geschaffen, Register effizient für statistische Zwecke nutzen zu können. Ähnlich stellt sich die Situation in den Niederlanden dar. Auch dort profitiert die Statistik davon, daß die Sozialversicherungsnummer in einer Reihe von Verwaltungsdateien gespeichert ist.

Für den Bereich der Wirtschaftsstatistiken sind derzeit Bestrebungen zur Einführung einer bundeseinheitlichen Unternehmensnummer im Gang. Die Fülle der hierbei auftretenden administrativen, rechtlichen und konzeptionellen Probleme lassen eine kurzfristige Realisierung dieses Ziels jedoch eher als unwahrscheinlich erscheinen.

Grundlagen der Zusammenführung von Datenbeständen ohne Identifikatoren

Während der Datenabgleich mit numerischen Identifikatoren DV-technisch vergleichsweise einfach durchzuführen ist, werfen Abgleiche ohne Identifikatoren ungleich schwierigere methodische Probleme auf. Dennoch ist es in vielen Bereichen auch außerhalb der amtlichen Statistik notwendig, Zusammenführungen ohne das Vorhandensein von Identifikatoren durchzuführen. Als Beispiele seien hier nur die Verschmelzung unterschiedlich strukturierter Kundendateien oder die Erstellung von Versanddateien im Bereich des Direct-Marketing genannt. Letztlich basiert aber auch das Suchen in großen Datenbanksystemen oder im Internet auf den nachfolgend beschriebenen Grundprinzipien.

Grundlage aller Verfahren zum Datenabgleich ohne numerische Identifikatoren bildet die Erkenntnis, daß jede Einheit (Person, Unternehmen etc.) bereits durch wenige Merkmale (sog. Sekundärschlüssel) weitgehend unverwechselbar charakterisiert werden kann und die Kombination der Merkmalsausprägungen fast schon einem Identifikator gleichkommt. Bei der Prüfung zweier Einheiten (Datensätze) auf Identität gelten folgende Grundsätze:

- Die zu vergleichenden Datensätze müssen gemeinsame Merkmale aufweisen.
- Je mehr gemeinsame Merkmale zum Vergleich zur Verfügung stehen, desto sicherer ist im allgemeinen das Ergebnis der Prüfung (identisch/nicht identisch).
- Je mehr Ausprägungen ein gemeinsames Merkmal annehmen kann, desto mehr kann es zu einer Identifikation beitragen.

Der zuletzt genannte Grundsatz bedarf einer Erläuterung. Betrachtet man den Datenabgleich als einen stufenweise verlaufenden Prozeß, dann wird durch jede Übereinstimmung von zwei Merkmalsausprägungen die Zahl der potentiell identischen Datensätze auf den Anteil der Häufigkeit der jeweiligen Merkmalsausprägung reduziert. Verwendet man bspw. bei einem Melderegisterabgleich das Merkmal Geschlecht, mit seinen zwei Ausprägungen männlich und weiblich, so wird hierdurch die Menge der in Frage kommenden Datensätze um etwa 50% reduziert. Einen deutlich höheren „Wirkungsgrad" haben Namensangaben. Familiennamen sind zwar nicht unverwechselbar und somit kein eigentlicher Identifikator, man denke nur an Meier, Schmidt oder Huber. Die Fülle der unterschiedlichen Namen (hier Merkmalsausprägungen) machen diese jedoch zu einem sehr wirkungsvollen Identifikationskriterium.

Die drei Verfahrensschritte des Datenabgleichs

In Anlehnung an Lenz[3)] läßt sich der Prozeß des Datenabgleichs, die Prüfung von Datensatzpaaren auf Identität, in drei Schritte unterteilen:

1. Bestimmung identifizierender Merkmale,
2. Vergleich der Datensatzpaare,
3. Klassifikation der Datensatzpaare in identisch/nicht identisch.

Die einzelnen Schritte des Datenabgleichs sollen im folgenden anhand einiger einfacher Beispiele in ihren Grundzügen erläutert werden. Aus Gründen der besseren Verständlichkeit wird auf eine mathematische Darstellung soweit wie möglich verzichtet.

Schritt 1: Bestimmung identifizierender Merkmale

Wie bereits erwähnt, ist das Vorhandensein mehrerer gleicher Merkmale in den zu vergleichenden Datenbeständen die Voraussetzung für einen Abgleich ohne numerische Identifikatoren. Die gemeinsamen Merkmale sind zu identifizieren und zu prüfen, ob sie inhaltlich weitgehend kongruent sind. Als problematisch können sich hierbei Merkmale erweisen, deren Ausprägungen sich im Zeitablauf verändern können, wie zum Beispiel das Merkmal Familienstand. Weisen die beiden zu vergleichenden Datenbestände Unterschiede hinsichtlich der Aktualität auf, besteht die Gefahr, daß es durch solch ein Kriterium zu fehlerhaften Identifikationen (sog. mismatches) kommt.

Schritt 2: Vergleich der Datensatzpaare

Dieser Schritt soll anhand des nachfolgenden Beispiels erläutert werden:

Beispiel:

Ein Datensatz a mit Personendaten soll mit einem Datenbestand B mit den Datensätzen $\{b_{1,\ldots,} b_n\}$ auf Identität geprüft werden. Als gemeinsame Merkmale $\{m_{1,\ldots,} m_4\}$ stehen zur Verfügung:

m_1: Vorname

m_2: Familienname

m_3: Geburtsdatum

m_4: Geschlecht (m/w)

Hinsichtlich der genannten Merkmale enthält der Datensatz a folgende Inhalte:

Datensatz	Vorname (m_1)	Familienname (m_2)	Geburtsdatum (m_3)	Geschlecht (m_4)
a	Peter	Müller	04.03.1967	m

Aus dem Bestand B kommen folgende Datensätze als potentiell identische Einheiten in Frage:

Datensatz	Vorname (m_1)	Familienname (m_2)	Geburtsdatum (m_3)	Geschlecht (m_4)
b_1	Peter	Miller	04.03.1967	m
b_2	Peter	Schmidt	04.03.1967	m

Unterstellt man bei einem maschinellen Vergleich das Trefferkriterium „volle Übereinstimmung der Merkmalsausprägungen“ so liefern beide gefundenen Datensätze das gleiche Ergebnis, nämlich Übereinstimmung in den drei Merkmalen Vorname, Geburtsdatum und Geschlecht mit der Folge, daß keine eindeutige Zuordnung erfolgen kann. Überließe man die Entscheidung, welcher der beiden Datensätze als identisch zu werten ist, hingegen einem Menschen, dann würde dieser sich aufgrund der Ähnlichkeit des Familiennamens für den Datensatz b_1 entscheiden.

Der Schritt 2, Vergleich der Datensatzpaare, kann sich folglich nicht nur auf die Feststellung „Übereinstimmung“ oder „keine Übereinstimmung“ beschränken, er beinhaltet auch die Messung einer teilweisen Übereinstimmung oder einer Ähnlichkeit der Merkmalswerte. Es stellt sich die Frage, mit welcher Berechtigung zwei Datensätze möglicherweise als identisch eingestuft werden, obwohl sie lediglich ähnlich sind. Der Grund liegt in der Datenqualität. Kein Datenbestand ist perfekt. Die Gründe sind vielfältig. Eine fast unvermeidbare Fehlerquelle stellt die Datenerfassung dar, sei sie manuell oder über Lesegeräte durchgeführt. Ebenso kaum vermeidbar sind Fehler bei der Datengewinnung. So kann bspw. eine Person im Melderegister mit dem Vornamen Hans-Jürgen registriert sein, bei einer Befragung aber den Rufnamen Hansi angegeben haben. Ferner spielt auch die Aktualität der Daten eine Rolle. Zum Beispiel kann ein Unternehmen, das in einem älteren Datenbestand noch mit Paul Meier OHG gespeichert ist, zwischenzeitlich die Rechtsform geändert haben und in dem aktuelleren Datenbestand als Meier AG verzeichnet sein.

Die Messung von Ähnlichkeiten bei Merkmalen mit wenigen Ausprägungen, wie bei dem Merkmal Geschlecht, ist weder sinnvoll noch durchführbar. Aber bereits beim Merkmal Geburtsdatum lassen sich Differenzierungen vornehmen, dahin gehend, daß eine Übereinstimmung nur des Geburtsjahres höher bewertet wird, als überhaupt keine Übereinstimmung.

Bei der überwiegenden Zahl von Datenabgleichen spielen Namen (Ortsnamen, Straßennamen, Familiennamen, Firmennamen etc.) eine entscheidende Rolle. Es ist daher nicht verwunderlich, daß Wissenschaft und Privatwirtschaft dem Thema „Messung von Namensähnlichkeiten“ besondere Aufmerksamkeit widmen. Zahlreiche Softwareunternehmen bieten Programme und Tools für den Abgleich von Dateien mit Namensangaben an. In der Literatur finden sich eine Reihe höchst unterschiedlicher Algorithmen zur Quantifizierung von Namensähnlichkeiten. Hierauf wird an späterer Stelle noch eingegangen.

Schritt 3: Klassifikation der Datensatzpaare in identisch/nicht identisch

Für einen systematischen maschinellen Abgleich von Dateien müssen feste Bewertungsregeln gefunden werden, ab wann zwei Datensätze als übereinstimmend zu gelten haben. Ziel ist die Formulierung einer Bewertungsregel, die zwei tatsächlich zueinander gehörende Datensätze mit größtmöglicher Sicherheit ermittelt. Eine solche Bewertungsregel besteht im allgemeinen aus drei Komponenten:

- den sog. **Vergleichsfunktionen $f_j(m_j)$** der gemeinsamen Merkmale m_j, mit denen Übereinstimmungen bzw. Ähnlichkeiten der Merkmalsausprägungen bewertet werden,
- der **Bewertungsfunktion $\lambda(f_j(m_j))$**, mit der für jedes Datensatzpaar die mit $f_j(m_j)$ bewerteten Merkmalsähnlichkeiten zu einem Übereinstimmungsmaß aggregiert werden, und
- der **Entscheidungsfunktion $\delta(\lambda)$**, welche die abgeglichenen Datensätze in identisch/nicht identisch klassifiziert.

Die einzelnen Komponenten sollen an einem Beispiel verdeutlicht werden. Nehmen wir wieder unseren Datensatz a aus dem vorherigen Beispiel und vergleichen ihn mit vier weiteren Datensätzen aus B:

Datensatz	Vorname (m_1)	Familienname (m_2)	Geburtsdatum (m_3)	Geschlecht (m_4)
b_3	Petra	Müller	04.03.1967	w
b_4	Hans-Peter	Müller	25.02.1967	m
b_5	Peter	Maurer	04.03.1967	m
b_6	Paul	Müller	12.08.1972	m

Bei der Entscheidung, welcher der vier Datensätze nun als identisch mit a zu werten ist, soll nach folgender Bewertungsregel vorgegangen werden:

– Bei dem Vergleich der Merkmale Vorname, Familienname und Geschlecht soll nur zwischen Treffer (=1) und kein Treffer (=0) unterschieden werden, beim Merkmal Geburtsdatum soll eine volle Übereinstimmung mit 1 und eine Übereinstimmung nur des Geburtsjahres mit 0,5 bewertet werden,

– es ist der Satz als identisch zu werten, der die höchste ungewichtete Treffersumme erzielt.

Bezeichnen wir die jeweils mit dem Datensatz a zu vergleichenden Datensätze aus B mit $\{b_1,\ldots, b_n\}$ dann gilt für die Vergleichsfunktionen $f_j(m_j)$ eines Datensatzpaares (a,b)

für die Merkmale m_1, m_2 und m_4:

$$f_j(m_j) = \begin{cases} 1, \text{ falls } m_j(a) = m_j(b) \\ 0, \text{ sonst} \end{cases} \qquad j=1,2,4$$

und für das Merkmal m_3:

$$f_3(m_3) = \begin{cases} 1, & \text{falls Geburtsdatum(a) = Geburtsdatum(b)} \\ 0{,}5, & \text{falls Geburtsdatum(a) } \neq \text{ Geburtsdatum(b),} \\ & \text{aber Geburtsjahr(a) = Geburtsjahr(b)} \\ 0, & \text{sonst} \end{cases}$$

Die Bewertungsfunktion $\lambda(f_j(m_j))$ eines Datensatzpaares läßt sich formulieren als:

$$\lambda(f_j(m_j)) = \sum_{j=1}^{4} f_j(m_j)$$

Für unser Beispiel ergibt sich somit folgende Bewertungstabelle:

Paar	$f_1(m_1)$	$f_2(m_2)$	$f_3(m_3)$	$f_4(m_4)$	λ
(a,b_3)	0	1	1	0	2
(a,b_4)	0	1	0,5	1	2,5
(a,b_5)	1	0	1	1	3
(a,b_6)	0	1	0	1	2

Die Entscheidungsfunktion

$$\delta(\lambda) = \begin{cases} \text{identisch, falls } \lambda = \max_i \lambda_i \\ \\ \text{nicht identisch, sonst} \end{cases}$$

mit λ_i (i=1,…,n) als den Bewertungen aller Datensatzpaare liefert uns Datensatz b_5 mit dem Maximalwert für λ als vermeintlich identisch zu a.

Unterstellen wir nun, daß das Merkmal m_2 Familienname zwingend übereinstimmen muß, dann lautet – bei sonst unveränderten Bewertungsregeln – nun die Bewertungsfunktion $\lambda(f_j(m_j))$:

$$\lambda(f_j(m_j)) = f_2(m_2) * (f_1(m_1) + f_3(m_3) + f_4(m_4))$$

Man erhält dann:

Paar	$f_1(m_1)$	$f_2(m_2)$	$f_3(m_3)$	$f_4(m_4)$	λ
(a,b_3)	0	1	1	0	1
(a,b_4)	0	1	0,5	1	1,5
(a,b_5)	1	0	1	1	0
(a,b_6)	0	1	0	1	1

Bei Anwendung dieser Entscheidungsregel wäre nun Datensatz b_4 als identisch zu a klassifiziert.

Welche Bewertungsregel gewährleistet die größtmögliche Sicherheit bei der Klassifikation der Datensätze in identisch/nicht identisch? Wie läßt sich eine solche Entscheidungsfunktion ermitteln?

Eine allgemeine Bewertungsregel existiert nicht. Jede Bewertungsregel, die eine weitgehend treffsichere Klassifikation gewährleisten soll, muß die spezielle Struktur, den Umfang und die Qualität der abzugleichenden Datenbestände berücksichtigen. Dies macht eine mehr oder weniger intensive Voruntersuchung der Datenbestände mittels einer Teststichprobe (testing sample) erforderlich. Bei diesem „Lernen“ der Klassifikationsregeln wird im allgemeinen wie folgt vorgegangen:

- Es wird eine Zufallsstichprobe von abzugleichenden Datensätzen gezogen.
- Zu jedem abzugleichenden Datensatz wird ein Satz potentiell identischer Datensätze ermittelt.
- Durch einen menschlichen Experten wird festgestellt, welcher der in Frage kommenden Datensätze tatsächlich der Identische ist.
- Aus dem Vergleich der möglichen Treffer mit dem tatsächlich identischen Satz werden dann die Vergleichsfunktionen der einzelnen Merkmale sowie deren Kombination zu einer Entscheidungsregel abgeleitet.

Für den letzten Schritt der Testuntersuchung können verschiedene methodische Werkzeuge, wie Regressions- oder Clusteranalysen, Entscheidungsbaumtechniken oder Simulationsmodelle herangezogen werden. Vielfach erfolgt das Aufstellen von Klassifikationsregeln aber auch anhand von Plausibilitätsüberlegungen, subjektiven Erfahrungen oder schlicht nach der Methode von Versuch und Irrtum.

In der Praxis liegen häufig bereits Informationen, Erfahrungen aus Abgleichen ähnlich strukturierter Datenbestände, vor. In diesen Fällen kann man auf bereits bewährte Entscheidungsregeln zurückgreifen, so daß sich dann die Testuntersuchung im wesentlichen auf das Formulieren der Vergleichsfunktionen, auch als Modellkalibrierung bezeichnet, beschränkt.

3. Zusammenführung von Datenbeständen im Rahmen des Zensustests

Das Grobkonzept des Zensustests

Die Notwendigkeit der Durchführung einer Volkszählung in Deutschland ist unbestritten. Während um die Jahrtausendwende nahezu alle Staaten der Europäischen Union sowie eine Vielzahl anderer Staaten (u.a. China, Indien, USA) Volkszählungen (Zensen) durchgeführt und damit neue Basisdaten über Gesellschaft, Staat und Wirtschaft gewonnen haben, datieren die „aktuellsten“ Volkszählungsergebnisse in Deutschland aus den Zählungen 1987 (alte Bundesländer) bzw. 1981 (neue Bundesländer). Bereits 1996 hatte die damalige Bundesregierung beschlossen, daß in Deutschland keine herkömmliche Volkszählung (Befragung aller Bürger durch Interviewer) mehr durchgeführt werden soll. Statt dessen wurde angeregt, die in vorhandenen Registern, vor allem in den Melderegistern der Gemeinden enthaltenen Daten zu nutzen. Die Verwendung von Verwaltungsregistern für statistische Zwecke und der hierdurch notwendige Einsatz neuer statistischer Verfahren birgt zweifellos gewisse Risiken und bedarf einer gründlichen Vorbereitung. Mit dem Gesetz zur Vorbereitung eines registergestützten Zensus hat der Gesetzgeber die amtliche Statistik beauftragt, zum Stichtag 05.12.2001 Testuntersuchungen zur Machbarkeit und Ausgestaltung eines künftigen registergestützten Zensus in Deutschland durchzuführen[4)].

Das Konzept für die Untersuchungen zu einem registergestützten Zensus (Zensustest) gliedert sich in drei Komponenten: Eine Stichprobe zur Feststellung von Mehrfachmeldungen in den Melderegistern („Mehrfachfallprüfung“), eine Stichprobe zur Feststellung von Über- und Untererfassungen in den Melderegistern („Registertests“) sowie eine multifunktionale Unterstichprobe. Mit dieser soll die im Rahmen des Zensustests bei Gebäude- und Wohnungseigentümern durchgeführte postalische Gebäude- und Wohnungszählung (GWZ[5)]) getestet, die Qualität der Dateien der Bundesanstalt für Arbeit untersucht und die Verfahren der Zusammenführung/Haushaltegenerierung („Verfahrenstests“) sowohl getestet als auch weiterentwickelt werden.

Die verschiedenen Testziele machen eine Reihe von Datenabgleichen erforderlich. Insbesondere bei den Untersuchungen zur Qualität der verwendeten Register und Verfahren liefert ein Ergebnisvergleich statistischer Aggregate nur unzureichende Erkenntnisse. So ist beispielsweise ein Vergleich von Einwohnerzahlen aus dem Melderegister und Direktbefragung in Folge der Saldierungseffekte von Über- und Untererfassungen im Melderegister nur wenig aussagekräftig. Erst der unmittelbare Vergleich der einzelnen Erhebungseinheiten erlaubt eine fundierte Qualitätseinschätzung.

Die wichtigsten Datenabgleiche im Rahmen des Zensustests sind:

1. Für den Test der Mehrfachfallprüfung in einem künftigen Zensus werden mittels einer Geburtstagsauswahl gezogene Stichproben aus den Melderegistern aller Gemeinden Deutschlands zu einem zentralen Bestand zusammengespielt. Zur Feststellung, ob Personen an mehr als einem Ort mit Hauptwohnsitz gemeldet sind, müssen durch ein entsprechendes Abgleichsverfahren identische Personen aufgespürt werden (Dublettensuche).

2. Zur Feststellung von Über- und Untererfassungen in den Melderegistern sowie der Qualität der gespeicherten demographischen Daten werden die an den Stichprobenadressen in den Registern verzeichneten Personendaten mit den durch Haushaltebefragungen vor Ort ermittelten Personendaten einzeln verglichen.
3. Die Feststellung und Beurteilung der Unterschiede zwischen den im Rahmen der GWZ postalisch bei den Gebäudeeigentümern/Verwaltern erhobenen und konventionell durch Haushaltebefragung gewonnenen Wohnungsdaten erfolgt durch Zusammenführung der beiden Erhebungsteile.
4. Zur Prüfung der Qualität der in den erwerbsstatistischen Dateien der Bundesanstalt für Arbeit enthaltenen Daten erfolgt eine Zusammenführung der Datei der sozialversicherungspflichtig Beschäftigten, der Arbeitslosendatei und der Datei der Teilnehmer an Maßnahmen der Fort- und Weiterbildung mit den Personendaten aus der Haushaltebefragung.
5. Mit dem Verfahren der Zusammenführung/Haushaltegenerierung sollen – sofern die Tests die Tauglichkeit des Verfahrens bestätigen – bei einem künftigen registergestützten Zensus haushalts- und wohnungsstatistische Daten gewonnen sowie mögliche Melderegisterfehler korrigiert werden. Neben der Haushaltsbildung anhand statistisch auswertbarer Merkmale des Melderegisters bildet die Zusammenführung von Personendaten aus dem Register mit den Wohnungsdaten der GWZ einen zentralen Bestandteil des Verfahrens.

In den Datenabgleichen 2 bis 5 stellen die in der Teststichprobe ausgewählten Adressen ein wichtiges Merkmal für den Abgleich der jeweiligen Datenbestände dar. Um methodisch schwierige und aufwendige Prüfungen von Orts- und Straßennamen zu vermeiden, wurde durch erhebungsorganisatorische Maßnahmen sichergestellt, daß in allen Erhebungsteilen ein eindeutiger numerischer Adressidentifikator bestehend aus dem amtlichen Gemeindeschlüssel und einer laufenden Adressnummer vorhanden ist. Die Datenabgleiche bestehen somit aus einem zweistufigen Prozeß: der Zusammenführung auf Adressebene mit Hilfe numerischer Identifikatoren und dann innerhalb der Adresse aus dem Einzelsatzabgleich der Personen- bzw. Wohnungsdaten mittels eines Verfahrens ohne Identifikator. Alle nachfolgenden Ausführungen beziehen sich auf die zuletzt genannte Stufe, den Abgleich auf Adressebene.

Die Zusammenführung von Melderegister- mit Wohnungsdaten

An das Verfahren zur Verknüpfung der Melderegister- mit den Wohnungsdaten im Verfahrensteil Zusammenführung/Haushaltegenerierung sind konzeptionell besondere Ansprüche zu stellen. Zum einen wegen der methodischen Schwierigkeiten, da hier im Gegensatz zu den anderen Datenabgleichen die als qualitativ besonders sicher geltenden Merkmale Geburtsdatum und Geschlecht nicht in beiden Datenbeständen zur Verfügung stehen und sich die Verknüpfungen im wesentlichen auf Namensmerkmale stützen müssen.[6)] Zum anderen ist zu berücksichtigen, daß dieser Verfahrensteil möglicherweise in einem künftigen registergestützten Zensus zur Anwendung kommen soll. In diesem Fall wären dann rund 37 Millionen Wohnungsdatensätze mit einem Gesamtbestand von ca. 90 Millionen Personendatensätzen abzugleichen. Hieraus ergeben sich zwei wesentliche Anforderungen an das Verfahren dieses Datenabgleichs:

- Angesichts der Datenvolumina und der Notwendigkeit, maschinell nicht identifizierbare Einheiten in einem personalintensiven und folglich kostenträchtigen Arbeitsgang manuell zusammenzuführen, muß das Verfahren eine möglichst hohe Trefferquote gewährleisten.
- Es sind aber auch die strengen Qualitätsmaßstäbe an Zensusergebnisse zu berücksichtigen. Jede fehlerhafte Zusammenführung eines Wohnungssatzes mit einem Personensatz kann zu Ergebnisverzerrungen führen. Es ist also eine möglichst hohe Treffsicherheit anzustreben.

Die Problematik besteht darin, daß beide Anforderungen in einem unmittelbaren Zielkonflikt zueinander stehen. Je „weicher“ die Bewertungsregeln für die Identifikation formuliert werden, desto größer wird zwar die Zahl der Zusammenführungen, desto größer ist aber auch die Gefahr, falsche Zuordnungen zu produzieren.

Ausgehend von diesen Randbedingungen wurde im Bayerischen Landesamt für Statistik und Datenverarbeitung ein Verfahren zur Verknüpfung von Wohnungssätzen aus der GWZ mit Personensätzen des Melderegisters entwickelt, das in modifizierter Form auch für die anderen Zusammenführungen im Rahmen des Zensustests zur Anwendung kommen soll. Dieses Verfahren wird im folgenden beschrieben.

Identifikation gemeinsamer Merkmale in Melderegister und GWZ (Schritt 1)

Um Wohnungsdaten mit den Personendaten des Melderegisters zusammenführen zu können, müssen diese kongruente Merkmale enthalten. Das bislang im Rahmen von Gebäude- und Wohnungszählungen übliche Erhebungsprogramm enthält kein Merkmal, mit dem sich eine Verknüpfung zu den Personendaten des Melderegisters herstellen ließe. Bei der Konzeption des Erhebungsprogramms der GWZ innerhalb des Zensustests war es daher erforderlich, über die Fragen zur Wohnung hinaus Fragen zum Namen und Einzugsdatum von maximal zwei Wohnungsinhabern, dies sind Hauptmieter oder Eigentümer, die ihre Wohnung selbst bewohnen, als Hilfsmerkmale für die Zusammenführung mit dem Melderegister aufzunehmen. In Schaubild 1 sind die entsprechend Schritt 1 zu vergleichenden Merkmale des Melderegisters und der GWZ aufgeführt.

Das deutsche Namensrecht läßt es zu, daß sich der Familienname einer Person entweder durch die Änderung des Familienstandes oder auf eigenen Antrag ändern kann. Aufgrund seiner Funktion als Verwaltungsregister enthält das Melderegister, wie aus Schaubild 1 ersichtlich, neben dem Familiennamen noch eine Reihe von Namensmerkmalen, die für den Abgleich herangezogen werden können. So könnte beispielsweise eine Person den Mietvertrag unter ihrem Geburtsnamen abgeschlossen haben und, da die Angaben der Eigentümer in der GWZ in der Regel aus den Mietverträgen stammen dürften, entsprechend in der GWZ verzeichnet sein, zwischenzeitlich aber durch Eheschließung einen anderen Familiennamen angenommen haben. In diesen Fällen wäre eine Identifikation nur über den Geburtsnamen möglich.

Die Zusammenführung von GWZ und Melderegister ist als zweistufiges Verfahren konzipiert. In der ersten Stufe

erfolgt der Datenabgleich zwischen den GWZ-Daten und den Datensätzen der an einer Adresse gemeldeten Personen. Führt dieser Abgleich zu keinem Ergebnis, so werden in einer zweiten Stufe die Verzeigerungen des Melderegisters verwendet. Unter Verzeigerungen versteht man die zu jeder Person in den Melderegistern eingetragenen Namen vorhandener Ehegatten, Kinder und gesetzlicher Vertreter. Diese Informationen können sehr nützlich sein, wenn der in der GWZ genannte Wohnungsinhaber selbst nicht an der untersuchten Adresse wohnt und folglich auch nicht identifiziert werden kann, aber eine mit ihm verzeigerte Person vorhanden ist (z.B. Kinder, für welche die Eltern den Mietvertrag abgeschlossen haben). Um in solchen Fällen eine Verknüpfung zwischen den Wohnungsdaten der GWZ und den Personendaten im Melderegister herstellen zu können, wird ein Abgleich anhand der Wohnungsinhabernamen der GWZ mit den im Melderegister in den Verzeigerungen enthaltenen Namen durchgeführt. Da dieser Abgleich methodisch weitgehend analog der Zusammenführung der Personendatensätze erfolgt, wird hierauf im folgenden nicht weiter eingegangen.

Schließlich war noch eine Vorbedingung für den Datenabgleich zu formulieren. Es ist äußerst unwahrscheinlich, daß Kinder im Alter unter 15 Jahren Inhaber einer Wohnung sind. Ausgehend von dem Stichtag des Zensustests, dem 05.12.2001, werden in dem Abgleich also nur Personen berücksichtigt, deren Geburtsdatum laut Melderegister vor dem 06.12.1986 liegt.

Vergleich der Datensatzpaare in Melderegister und GWZ (Schritt 2)

Wie bereits erwähnt, spielen Namensübereinstimmungen für die Verknüpfung der Einzeldaten aus GWZ und Melderegister die zentrale Rolle. Da davon auszugehen war, daß beide Datenquellen hinsichtlich der Namensschreibweisen gewisse qualitative Mängel aufweisen, mußten bei dem Vergleich der Datensatzpaare auch Namensähnlichkeiten Berücksichtigung finden. In der Literatur finden sich eine Reihe von DV-technisch umsetzbaren Verfahren zur Messung bzw. Bewertung von Namensähnlichkeiten. Bevor nachfolgend auf eine im Bayerischen Landesamt für Statistik und Datenverarbeitung speziell für die Anforderungen des Zensustests entwickelte Methode eingegangen wird, sollen die zwei bekanntesten Algorithmen zur Bewertung von Namensähnlichkeiten, der „Soundex-Algorithmus“ und der „fehlertolerante Vergleich“ vorgestellt werden.

Der Soundex – Algorithmus

Ein bekanntes Verfahren für einen Namensabgleich ist der sog. Soundex – Algorithmus, der von Donald Knuth entwickelt wurde und eine weit verbreitete Anwendung erreicht hat. Grundlage des Algorithmus ist eine Codierung der zu vergleichenden Namen. Der Soundex-Code wird folgendermaßen gebildet[7]:

1. Umlaute werden zu AE, OE und UE
2. Der erste Buchstabe des Namens wird vorgemerkt
3. Alle Buchstaben des Namens nach dem Anfangsbuchstaben werden folgendermaßen codiert:

   A, E, I, O, U, W, Y, H = 0

   B, P, F, V = 1

   C, S, K, G, J, Q, X, Z = 2

   D, T = 3

   L = 4

   M, N = 5

   R = 6
4. Hintereinander stehende gleiche Ziffern eines Codes werden jeweils zu einer Ziffer zusammengefaßt.
5. Die erste Ziffer sowie sämtliche Nullen des Codes werden gestrichen
6. Der vorgemerkte Anfangsbuchstabe wird dem Code vorangestellt. Ab der vierten Ziffer werden sämtliche Ziffern gelöscht.
7. Verbleiben nach den sechs genannten Schritten weniger als drei Ziffern so wird der Code bis einschließlich zur dritten Ziffer mit Nullen aufgefüllt.

Der fertige Soundex – Code besteht also immer aus dem Anfangsbuchstaben und drei Ziffern.

Beispiel: Müller Karl und Miller Carl

	Nachname	**Vorname**	**Nachname**	**Vorname**
Ausgangslage	Müller	Karl	Miller	Carl
Schritt 1	Mueller	Karl	Miller	Carl
Schritt 2	M	K	M	C
Schritt 3	5004406	2064	504406	2064
Schritt 4	50406	2064	50406	2064
Schritt 5	46	64	46	64
Schritt 6	M46	K64	M46	C64
Schritt 7	M460	K640	M460	C640

Sind die Soundex – Codes zweier Namen gleich, so werden die Namen als übereinstimmend betrachtet. Im obigen Beispiel würde also der alleinige Vergleich von Müller und Miller zu einem Treffer führen, nicht jedoch der Vergleich von Müller Karl und Miller Carl.

Das Beispiel zeigt eine gewisse Anfälligkeit des Verfahrens, wenn zwei Namen unterschiedliche Anfangsbuchstaben aufweisen. Dies kann jedoch durch Schreib- oder Übertragungsfehler ähnlich ausgesprochener Namen relativ leicht geschehen, beispielsweise bei Prechtl und Brechtl oder Wladimir und Vladimir.

Die relativ simple, auf die englische Aussprache ausgelegte Codierung bedingt auch, daß Namen, die deutlich voneinander abweichen, denselben Code aufweisen und damit als identisch bewertet werden würden. Beispiele hierfür sind:

Euler und Ellery = E460

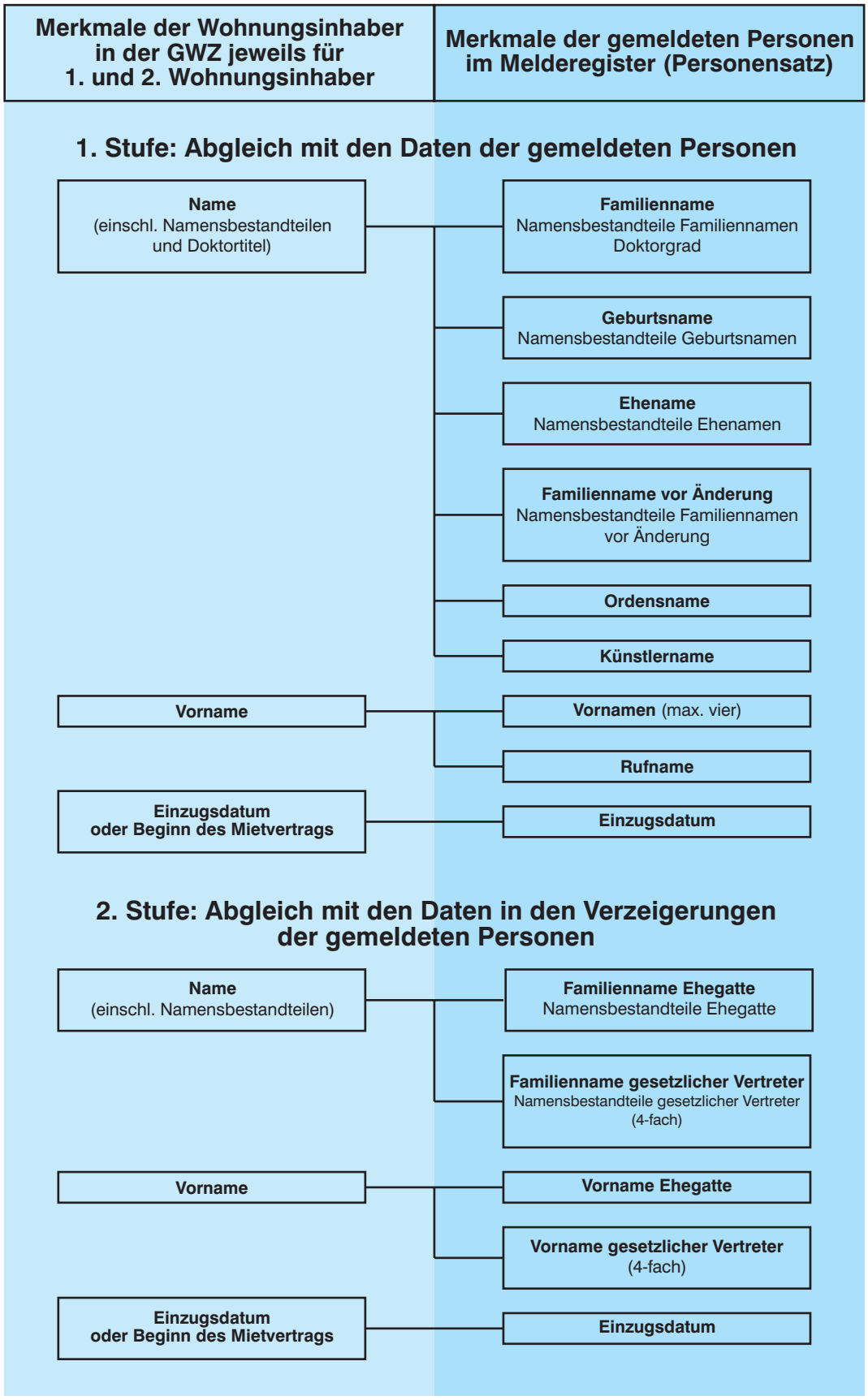
Gauss und Ghosh = G200

Knuth und Kant = K530

Lloyd und Ladd = L300

Lukasiewicz und Lissajous = L222

Eulert hätte hingegen den Code E463 und Kanter den Code K536. Wäre der Name des Wohnungsinhabers Euler und stünden im Melderegister die Namen Eulert und Ellery, so würde der Soundex – Algorithmus den Namen Euler mit Ellery zusammenführen und nicht mit dem viel näher liegenden Eulert. Zudem bestünde auch das Problem von Mehrfachtreffern: Wenn beispielsweise der Wohnungsinhaber Latt heißt und im Melderegister ein Lloyd und ein Ladd verzeichnet sind, so kann der Soundex – Algorithmus keine eindeutige Zuordnung vornehmen. Schon anhand dieser sehr einfachen Beispiele wird deutlich, daß der Soundex – Algorithmus für die speziel-



CD 36002/Sg25 Bayerisches Landesamt für Statistik und Datenverarbeitung

Schaubild 1

len Belange des Namensabgleichs von Daten aus den Melderegistern und der Gebäude- und Wohnungszählung nicht geeignet ist.

Der fehlertolerante Vergleich (Trigramm – Ähnlichkeitsmaß)

Beim fehlertoleranten Vergleich[8)] erfolgt keine Codierung der Namen. Der Vergleich erfolgt anhand des Grades an Übereinstimmung der in den beiden Namen enthaltenen Trigramme. Unter einem Trigramm werden drei aufeinanderfolgende Buchstaben eines Namens verstanden. Beispielsweise enthält der Name Meier die Trigramme MEI, EIE und IER. Die Übereinstimmung zweier Namen wird anhand der Formel

$s = 2c/(n+m)$

gemessen, wobei s den Grad der Übereinstimmung, c die Zahl der gemeinsamen Trigramme und n bzw. m die gesamte Zeichenlänge der beiden Namen angibt. Die beiden Namen Maier und Meier hätten also beispielsweise eine Übereinstimmung von 0,2. Für einen Vergleich von Namen müßten Schwellenwerte für s festgelegt werden, ab deren Überschreitung zwei Namen als zusammengehörig betrachtet werden können.

Ein Nachteil des Verfahrens ist, daß nur die Zahl der gemeinsamen Trigramme in die Bewertung eingeht, das Ausmaß der Abweichungen jedoch nur eine verhältnismäßig niedrige Rolle spielt. So haben die Namen Meier und Mejer keine gemeinsamen Trigramme, s wäre also in diesem Fall gleich 0. Meier und Forrestier haben hingegen das Trigramm IER gemeinsam, s wäre in diesem Fall gleich 0,13. Eine naive Anwendung des Trigramm – Ähnlichkeitsmaßes würde also eher Meier und Forrestier als Meier und Mejer zusammenführen. Ein solcher Extremfall könnte durch die Wahl eines geeigneten Schwellenwerts von s verhindert werden, der einen Treffer erst bei einer verhältnismäßig hohen Übereinstimmung zuläßt. Allerdings bliebe auch dann vor allem bei kurzen Namen ein gewisses Risiko, ähnliche Namen nicht zu finden, wie das Beispiel Meier und Mejer verdeutlicht.

Gegenüber dem Soundex – Algorithmus hat das Trigramm – Ähnlichkeitsmaß den Vorteil, daß es eine Quantifizierung der Ähnlichkeit zweier Namen und damit auch eine Reihung möglicher Treffer zuläßt, während es beim Soundex – Algorithmus lediglich eine Übereinstimmung oder Nicht – Übereinstimmung gibt. Der entscheidende Nachteil ist die Abhängigkeit von den gemeinsamen Trigrammen, die vor allem bei sehr kurzen Namen zu fehlerhaften Zuordnungen führen kann. Gerade solche Fehler müssen jedoch bei der Zusammenführung zweier Datenquellen unbedingt vermieden werden.

Das Phonetikmodul

Obwohl es sich nicht um ein phonetisches Verfahren handelt, hat sich für das im Bayerischen Landesamt für Statistik und Datenverarbeitung entwickelte Verfahren zur Bewertung von Namensähnlichkeiten fälschlicherweise der Begriff Phonetikmodul eingebürgert. Unter phonetischen Verfahren werden Techniken verstanden, bei denen sich die Messung der Ähnlichkeit wie im oben beschriebenen Soundex-Algorithmus im wesentlichen auf einen Vergleich des sprachlichen Klangs der jeweiligen Namen stützt. Diese sind folglich vor allem geeignet für den Abgleich von Datenbeständen, bei denen die Namen aufgrund akustischer Übermittlungen (z.B. durch Interviews) erfaßt wurden. Anders stellt sich die Situation beim Zensustest dar. Hier erfolgt die Datenerfassung anhand schriftlicher Belege. Im Melderegister sind dies Meldescheine, Standesamtsmitteilungen etc., bei der GWZ die ausgefüllten Fragebogen. Deshalb kann in den meisten Fällen davon ausgegangen werden, daß Unterschiede (identischer) Namen zum Teil aus visuell oder manuell bedingten Erfassungsfehlern resultieren.

Ein häufiger Fehler bei der Erfassung handschriftlicher Belege ist die Verwechslung von „u“ und „n“. Es sei zum Beispiel eine Person in der GWZ mit dem Namen „Bandisch“ erfaßt worden, der im Melderegister verzeichnete Name lautet hingegen „Baudisch“. Phonetisch betrachtet ist die Ähnlichkeit zwischen beiden Namen nicht sehr stark ausgeprägt. Vergleicht man jedoch die einzelnen Buchstaben und ihre Reihenfolge, kann von einer sehr deutlichen Übereinstimmung, nämlich sieben von acht Zeichen, gesprochen werden.

Damit ist bereits das Grundprinzip des sog. Phonetikmoduls angesprochen. Tatsächlich handelt es sich um ein Verfahren, das die Ähnlichkeit von Namen anhand der Übereinstimmungen von einzelnen Buchstaben bzw. Buchstabenfolgen quantifiziert (Zeichenkettenvergleich). Nach einer Vorbereitungsstufe (Umsetzung in Großbuchstaben, Entfernen von Sonderzeichen etc.) wird ein String (hier der Wohnungsinhabername) in einzelne Zeichen oder Zeichenketten („Teilstrings“) zerlegt und dann überprüft, ob diese Teile im Vergleichsstring (hier ein Melderegistername) enthalten sind. Der Grad der Ähnlichkeit zweier Strings kommt in der Bewertungsziffer zum Ausdruck. Diese ist definiert als die Summe der Längen der gemeinsamen Teilstrings, dividiert durch die Länge des längeren der beiden Strings. Per definitionem muß die Bewertungsziffer zwischen 0 und 1 liegen.

In dem vorherigen Beispiel ergibt sich dann:

B A N **D I S C H**

B A U **D I S C H**

Länge der gemeinsamen Teilstrings: 2+5

Länge des längeren Namens: 8

Bewertung: 7/8 = 0,88

Es stellt sich zunächst die Frage nach der Mindestteilstringlänge, also der Länge, die ein Teilstring aufweisen muß, um als Treffer akzeptiert zu werden. Ist z.B. die Mindestteilstringlänge = 2, gehen einzelne Buchstaben nicht in die Bewertung ein. Eine Übereinstimmung wird erst dann berücksichtigt, wenn mindestens zwei aufeinanderfolgende Buchstaben übereinstimmen. Grundsätzlich ist die Mindestteilstringlänge in der programmtechnischen Umsetzung des Verfahrens beliebig einstellbar. Es hat sich allerdings durch Sensitivitätsanalysen an Testdaten gezeigt, daß es sinnvoll ist, die Mindestlänge von der Länge der zu vergleichenden Strings abhängig zu machen. Das günstigste Verhältnis zwischen Trefferquote und Treffersicherheit wurde in dem Testmaterial mit einer Mindestteilstringlänge = 1 bei Namen mit weniger als sieben Buchstaben und einer Mindestteilstringlänge = 2 bei Namen mit einer Länge von sieben oder mehr Buchstaben erzielt.

Meist ist es wenig sinnvoll, die Buchstaben zweier Strings positionsgenau miteinander abzugleichen, also die erste Stelle von Name 1 mit der ersten Stelle von Name 2, die zweite Stelle von Name 1 mit der zweiten Stelle von Name 2 usw.; da bei unterschiedlichen

Schaubild 2. Beispiel zum Prozeß des Zeichenkettenvergleichs (Mindestteilstringlänge = 2, „links„ = 2, „rechts“ = 2)

Schritt	Gesuchter String			Vergleichsstring			Treffer ja/nein	Trefferanzahl	Trefferlänge
	DRAKOMENA			DRAOMINA					
	Reststring	gesuchter Teilstring	Reststring	Reststring	Suchraum	Reststring			
1		**DR**	AKOMENA		**DR**AO	MINA	Ja	1	2
2		**DRA**	KOMENA		**DRA**OM	INA	Ja	1	3
3		**DRAK**	OMENA		DRAOMI	NA	Nein	–	–
4	DRA	**KO**	MENA	D	RAOMIN	A	Nein	–	–
5	DRAK	**OM**	ENA	DR	A**OM**INA		Ja	2	2
6	DRAK	**OME**	NA	DR	AOMINA		Nein	–	–
7	DRAKOM	**EN**	A	DRAO	MINA		Nein	–	–
8	DRAKOME	**NA**		DRAOM	I**NA**		Ja	3	2

Schreibweisen, z.B. aufgrund eines Erfassungsfehlers, ein und derselbe Buchstabe nicht mehr an derselben Position im Vergleichsstring zu finden sein wird. So würde beispielsweise bei einem Vergleich der beiden Vornamen ARON und AARON einzig der Buchstabe „A“ als gemeinsames Element erkannt werden, was die tatsächliche Ähnlichkeit kaum noch widerspiegelt.

Deshalb wird ein Suchraum, das heißt ein Intervall, innerhalb dessen der Vergleich der Teilstrings stattfinden soll, festgelegt. Der Suchraum im Vergleichsstring umfaßt jetzt nicht mehr ausschließlich die Positionen der Buchstaben des gesuchten Teilstrings im ursprünglichen Namen, sondern wird links und rechts um eine vorgegebene Anzahl von Positionen ausgeweitet. Die Erweiterung der Positionen wird durch die Parameter „links“ (= vor dem Teilstring) und „rechts“ (= nach dem Teilstring) festgelegt. Der so gebildete Suchraum kann jedoch keine Positionen außerhalb des Vergleichsstrings umfassen. Befindet sich der gesuchte Teilstring am Beginn bzw. am Ende des Vergleichsstrings, wird der Suchraum deshalb an der entsprechenden Seite verkürzt.

In dem folgenden Beispiel soll der Teilstring „**OM**“ aus dem Namen „DRAKOMENA“ im Vergleichsstring „DRAOMINA“ gefunden werden. Die Parameter für „links“ und „rechts“ haben jeweils den Wert 2.

Position:	1	2	3	4	5	6	7	8	9
Name:	D	R	A	K	**O**	**M**	E	N	A
Vergleichsstring:	D	R	A	**O**	**M**	I	N	A	

Bei der Suche nach dem Teilstring „**OM**“, den Stellen 5 und 6 des Strings „DRAKOMENA“, werden nicht allein die Stellen 5 und 6 (“MI“) im Vergleichsstring „DRAOMINA“ betrachtet, sondern der Suchraum wird jeweils nach links um zwei Stellen bis einschließlich Stelle 3 und nach rechts um zwei Stellen bis einschließlich Stelle 8 ausgedehnt. Dadurch ergeben sich für den Suchraum im Vergleichsstring „DRAOMINA“ die Intervallgrenzen [3; 8]. Der Teilstring „OM“ wird an den Stellen 4 und 5 gefunden und als Übereinstimmung gewertet.

Je größer der Suchraum, also die Parameter „links“ und „rechts“ gewählt werden, desto mehr Übereinstimmungen werden gefunden. Ein zu großer Suchraum bedingt aber die Gefahr von Fehlern, wie der Vergleich der kurzen Namen BAER und RABE bei einer Mindestteilstringlänge von 1 zeigt. Ohne eine Beschränkung der Intervallgrenzen würden beide Wörter irrtümlicherweise eine volle Übereinstimmung aufweisen.

Die Parameter „Mindestteilstringlänge“ und „Suchraum“ stehen in unmittelbarer Abhängigkeit. Grundsätzlich gilt für einen effizienten Vergleich: Je länger die Mindestteilstringlänge, desto größer kann der Suchraum gewählt werden und umgekehrt. Bei den Testuntersuchungen haben sich für die Parameter „links“ und „rechts“, welche die Größe des Suchraums bestimmen, sowie für die Mindestteilstringlänge der Teilstrings einige Grundeinstellungen etabliert.

Für Namen mit sieben oder mehr Zeichen gilt in der Regel:

Mindestteilstringlänge = 2

„links“ = 2

„rechts“ = 2

Für kurze Namen (beide Strings, d.h. GWZ-Name und Melderegistername haben jeweils weniger als sieben Buchstaben) haben sich folgende Werte als tragfähig erwiesen:

– Länge(GWZ-Name) < Länge(MR-Name):
  Mindestteilstringlänge = 1, „links“ = 0, „rechts“ = 1
– Länge(GWZ-Name) > Länge(MR-Name):
  Mindestteilstringlänge = 1, „links“ = 1, „rechts“ = 0
– Länge(GWZ-Name) = Länge(MR-Name):
  Mindestteilstringlänge = 1, „links“ = 0, „rechts“ = 0

Am Beispiel in Schaubild 2 wird der Prozeß des Zeichenkettenvergleichs in Abhängigkeit der Parameter nochmals verdeutlicht, Schaubild 3 zeigt einige Beispiele für quantifizierte Ähnlichkeiten.

Schaubild 3. Beispiele zur Bewertung des Zeichenkettenvergleichs (Mindestteilstringlänge = 2, „links„ = 2, „rechts“ = 2)

Gesuchter String	Vergleichsstring	Länge der gemeinsamen Teilstrings	Länge des längeren Namens	Bewertungsziffer
DRA**KOM**E**NA**	**DRAOM**I**NA**	3+2+2	9	7/9 = 0,78
RIEKI	**RILKI**	2+2	5	4/5 = 0,80
ATANASSION**I**	**ATANASIO**U	6+2	11	8/11 = 0,73
LIESKOVSKY	**LIESZKOVSZKY**	4+4+2	12	10/12 = 0,83
JEANETTE	**JEAN**N**ETTE**	4+4	9	8/9 = 0,89

Die große Stärke des Verfahrens ist die hohe Flexibilität. Durch Variation der Mindestteilstringlänge sowie der Parameter „links“ und „rechts“ jeweils in Verbindung mit den jeweiligen Namenslängen kann das Verfahren auf Spezialfälle kalibriert werden, wie das folgende Beispiel zeigt:

Es soll die Ähnlichkeit der beiden Vornamen GRETE und ANNEGRET bewertet werden. Bei einer Mindestteilstringlänge von 2 führt der Vergleich zu einer Bewertung von 0, da der Teilstring „GRET“ in den beiden Vornamen an zu weit voneinander entfernt liegenden Stellen positioniert ist. Um derartige offensichtlich zusammengehörige Sätze dennoch zusammenführen zu können, muß der zulässige Suchraum deutlich erweitert werden (z.B. „links“ = 7, „rechts“ = 7). Um sicherzustellen, daß die richtigen Sätze zusammengeführt werden, wird im Gegenzug die geforderte Mindestteilstringlänge in direkter Abhängigkeit von der Länge des kürzeren Vornamens stark heraufgesetzt (und zwar auf Länge(GWZ-Vorname)-1 bzw. Länge(Melderegister-Vorname)-1). In unserem Beispiel beträgt die Mindestteilstringlänge somit 4 (= Länge(GRETE) – 1), d.h. der kürzere Vorname muß sich nahezu vollständig im längeren Vornamen finden lassen.

Konzept der Zusammenführung (Schritt 3)

Wie in Schritt 3 „Klassifikation der Datensatzpaare in identisch/nicht identisch“ erläutert, müssen Regeln für die Bewertung von potentiellen Übereinstimmungen sowie für die Entscheidung, ab wann zwei Datensätze als übereinstimmend zu werten sind, aufgestellt werden. Theoretisch bedeutet der Vergleich zweier Datenbestände, daß ein paarweiser Vergleich zwischen jedem Satz des einen Datenbestandes mit jedem Satz des anderen Datenbestandes durchzuführen wäre. Bezeichnen wir den GWZ-Bestand als A mit den Datensätzen $\{a_1,\ldots, a_n\}$ und den Melderegisterbestand mit B $\{b_1,\ldots, b_m\}$, dann ergibt sich die Menge der zu überprüfenden Datensatzpaare C aus dem Kreuzprodukt von A und B:

$C = A \quad B$ mit $C = \{(a_1,b_1);(a_1,b_2);\ldots;(a_n,b_m)\}$

Bei einem vollständigen Vergleich wären also je Adresse n*m Datensatzvergleiche erforderlich. Auf diese Menge der Vergleichsergebnisse müßte dann eine entsprechende Entscheidungsfunktion δ angewendet werden, die ermittelt, welche Datensatzpaare zusammengehören. Diese Vorgehensweise wäre wenig praktikabel. Ist beispielsweise ein Datensatzpaar hinsichtlich der gemeinsamen Merkmale völlig übereinstimmend, würden beide Datensätze dennoch mit allen anderen Datensätzen verglichen werden, obwohl dieses Ergebnis nicht mehr zu übertreffen ist.

Deutlich weniger Vergleichsoperationen und somit auch kürzere Rechnerlaufzeiten entstehen bei folgendem Vorgehen: Es wird zunächst der erste Datensatz (GWZ) mit allen Datensätzen des Melderegisters verglichen und aufgrund der m Vergleichsergebnisse eine Klassifikation in identisch/nicht identisch vorgenommen. Anschließend wird diese Prozedur mit den weiteren Datensätzen der GWZ wiederholt. Da ein als identisch klassifizierter Datensatz des Melderegisters nun nicht mehr in den Vergleich einbezogen wird, findet eine schrittweise Reduktion der Vergleichsoperationen (sukzessive Massenreduktion) statt[9)].

Wie das folgende stark vereinfachte Beispiel zeigt, beinhaltet diese Vorgehensweise allerdings ein gravierendes methodisches Problem.

GWZ

Lfd. Nr.	Name	Vorname
1	Müller	Petra
2	Müller	Peter

Melderegister

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

Führt man beginnend mit dem Datensatz Nr. 1 des GWZ-Materials den Vergleich mit den Datensätzen des Melderegisters durch, so kann es passieren, daß aufgrund der vollen Übereinstimmung des Namens „Müller“ und der starken Ähnlichkeit der Vornamen „Petra“ und „Peter“ der Datensatz Nr. 3 des Melderegisters fälschlicherweise als zu Datensatz Nr. 1 der GWZ identisch klassifiziert wird. Da der Datensatz Nr. 3 des Melderegisters bei den nächsten Vergleichen nicht mehr betrachtet wird, kann die tatsächliche Identität zwischen den Datensätzen Nr. 2 der GWZ und Nr. 3 des Melderegisters nicht mehr festgestellt werden.

Das hierarchische Stufenmodell

Das Beispiel zeigt, daß durch jede ausgeführte Verknüpfung die spätere Feststellung einer noch größeren Übereinstimmung mit einem anderen Datensatz unmöglich wird. Diese Schwäche des Verfahrens kann durch die Verwendung mehrerer hierarchisch abgestufter Bewertungsregeln weitgehend vermieden werden. Hierbei wird die oben beschriebene Prozedur zunächst mit einer sehr restriktiven Bewertungsregel, also beispielsweise volle Übereinstimmung aller gemeinsamer Merkmale, durchlaufen. Für alle weiteren Stufen, deren Bewertungsregeln verglichen mit der ersten Stufe zunehmend „weicher“ werden – z.B. Nachnamen völlig identisch, aber Vornamen nur zu einem gewissen Grad ähnlich -, müssen jeweils nur noch jene GWZ- und Melderegistersätze miteinander verglichen werden, die noch nicht in einer vorhergehenden Stufe verknüpft wurden. Nach diesem Prinzip wurde das Verfahren zur Verknüpfung der Wohnungsdaten mit Melderegisterdaten konzipiert.

Die Aufstellung der hierarchischen Bewertungsregeln zur Klassifikation der Datensatzpaare in identisch/nicht identisch wurde auf der Grundlage einer Auswertung umfangreicher Testdaten vorgenommen. Für die einzelnen Komponenten gelten folgende Grundsätze.

Für die Vergleichsfunktionen $f_j(m_j)$ der gemeinsamen Merkmale m_j, mit denen Übereinstimmungen bzw. Ähnlichkeiten der Merkmalsausprägungen bewertet werden, gilt nach Prioritäten geordnet:

- Die volle Übereinstimmung zweier Namen erhält die höchste Bewertung von 1.
- Namensähnlichkeiten werden mit dem oben beschriebenen Phonetikmodul gemessen, das hierbei erzielte Ähnlichkeitsmaß im Intervall [0,1] stellt auch die Bewertung der Ähnlichkeit dar. Je nach Bewertungsstufe liegt der Schwellenwert der Ähnlichkeit bei 0,75 bzw. 0,5, d.h. darunter liegende Ähnlichkeitswerte werden mit 0 bewertet.
- Wird eine unter dem Schwellenwert liegende Ähnlichkeit festgestellt, erfolgt bei den Nachnamen eine Prüfung auf Mehrfachnamen (z.B. Müller-Lüdenscheid). Volle Übereinstimmung oder Ähnlichkeit eines Namensteils gehen als teilweise Übereinstimmung in die Bewertung ein.
- Der Vergleich zwischen den Feldern Einzugsdatum in den Datenbeständen erfolgt anhand des Absolutbetrages der Differenz zwischen Einzugsdatum laut GWZ und Einzugsdatum laut Melderegister.

Hinsichtlich der Bewertungsfunktionen $\lambda(f_j(m_j))$, mit der für jedes Datensatzpaar die mit $f_j(m_j)$ bewerteten Merkmalsähnlichkeiten zu einem Übereinstimmungsmaß aggregiert werden, wurden folgende Grundsätze berücksichtigt:

- Beim Vergleich eines Datensatzpaares müssen Name und Vorname zumindest teilweise Übereinstimmungen bzw. Ähnlichkeiten ($f_j(m_j) > 0$) aufweisen, um überhaupt als potentiell identische Datensätze in Frage zu kommen. Übereinstimmungen nur bei einem Namensmerkmal allein sind nicht ausreichend.
- Das Merkmal Einzugsdatum hat sich im Testmaterial nur bedingt als tauglich erwiesen und fließt daher nicht in die Bewertungsfunktion ein.
- Die hierarchische Anordnung der Bewertungsregeln gewährleistet implizit, daß Übereinstimmungen von (Nach-)Namen aufgrund der höheren Variabilität der Ausprägungen ein höheres Gewicht zukommt.
- Die Bewertung des Vergleichs zwischen dem Namen (GWZ) und den verschiedenen Namensfeldern (einschl. Namensbestandteilen) des Melderegisters erfolgt nach folgenden Prioriäten:

  Familienname → Geburtsname → Ehename → Familienname vor Änderung → Künstlername → Ordensname.
- Für die Bewertung des Vergleichs zwischen dem Vornamen (GWZ) und den Vornamensmerkmalen des Melderegisters gelten allgemein folgende Abstufungen:
  - Übereinstimmung Vorname (GWZ) mit allen im Feld Vornamen des Melderegisters aufgeführten Vornamen
  - Übereinstimmung Vorname (GWZ) mit Rufnamen (Melderegister)
  - Übereinstimmung Vorname (GWZ) mit einem der im Feld Vornamen des Melderegisters aufgeführten Vornamen

Schließlich gilt für die Entscheidungsfunktion $\delta(\lambda)$, welche die abgeglichenen Datensatze in identisch/nicht identisch klassifiziert:

- Das Datensatzpaar mit der höchsten Bewertung wird als identisch klassifiziert. Hierbei erhält der Datensatz des Melderegisters die Wohnungsnummer aus dem GWZ-Bestand und der Wohnungssatz die interne Ordnungsnummer der Person aus dem Melderegisterdatensatz.
- Sind zwei oder mehr Datensatzpaare gleich bewertet (Indifferenz), so erfolgt die Entscheidung anhand des Kriteriums Einzugsdatum.

Die Hauptstufen des Zusammenführungskonzepts

Auf der Basis dieser Grundsätze wurde ein in vier Hauptstufen untergliedertes 43 Bewertungsregeln (Unterstufen) umfassendes hierarchisches Bewertungssystem entwickelt, das nachfolgend grob skizziert wird.

Voraussetzung für die **erste Hauptstufe** ist die völlige Übereinstimmung des Namens aus der GWZ mit den verschiedenen Nachnamen des Melderegisters. Zunächst wird völlige Identität des Namens aus der GWZ mit dem Familiennamen aus dem Melderegister (unter Berücksichtigung von Namensbestandteilen und Doktorgrad) postuliert und in verschiedenen Unterstufen die Vornamen auf Übereinstimmung bzw. Ähnlichkeiten überprüft. Entsprechend wird dann mit Geburtsnamen, Ehenamen und Familiennamen vor Änderung verfahren.

Im Aufbau weitgehend analog ist die **zweite Hauptstufe**. Der wesentliche Unterschied besteht darin, daß nun keine vollständige Übereinstimmung, sondern nur noch eine hinreichend große Ähnlichkeit zwischen den Nachnamen vorausgesetzt wird.

Die nach den Hauptstufen 1 und 2 noch verbleibenden unverknüpften Datensätze gelangen in die Unterstufen der **dritten Hauptstufe**. Dort werden Mehrfachnamen (nur Nachnamen) auf teilweise Übereinstimmung geprüft und in Verbindung mit den Vornamensprüfungen bewertet. Seitens des Melderegisters sind in dieser Hauptstufe nur die Familiennamen einbezogen.

Die **vierte Hauptstufe** umfaßt Sonderformen des Datenabgleichs. Zunächst finden Prüfungen auf Vertauschungen von Vor- und Familiennamen statt. Insbesondere asiatische Namen weisen häufig mehrere Namensteile auf, die nicht eindeutig in Vor- und Nachnamen unterscheidbar sind. Durch Verkettungen und Vertauschungen von kompletten Vor- und Nachnamen werden eventuell vorhandene Übereinstimmungen aufgedeckt und bewertet. In den letzten Unterstufen werden schließlich potentielle Übereinstimmungen hinsichtlich Künstler- und Ordensnamen überprüft.

Die Funktionsweise der Unterstufen – ein Beispiel

Da eine ausführliche Darstellung aller Unterstufen den Rahmen dieses Beitrages sprengen würde, wird nachfolgend eine ausgewählte Unterstufe exemplarisch vorgestellt. Als Beispiel wurde die Unterstufe 215 gewählt. Diese zählt zur Hauptstufe 2, d.h. es wird „nur“ eine hinreichende Ähnlichkeit der Nachnamen vorausgesetzt. Weiterhin werden, nachdem in den vorherigen Unterstufen die Vornamen auf vollständige Übereinstimmung geprüft wurden, nun die Vornamen auf Ähnlichkeit untersucht.

Es sei nun (a,b) ein Datensatzpaar aus der Menge aller möglichen Datensatzpaare eines noch unverknüpften Datensatzes der GWZ und den noch unverknüpften Datensätzen des Melderegisters. Weiterhin bezeichnet p(N) die in dem Datensatzpaar (a,b) mit dem Phonetikmodul festgestellte Ähnlichkeit hinsichtlich des gemeinsamen Merkmals „Nachname“ (Name-GWZ und Familienname-Melderegister) und p(V) die des gemeinsamen Merkmals „Vorname“ (Vorname-GWZ und Rufname-Melderegister).

Eine mögliche Zusammenführung in der Unterstufe 215 soll nur bei Datensatzpaaren erfolgen, bei denen eine phonetische Übereinstimmung beider Merkmale von mindestens 0,5 vorliegt. Für die Vergleichsfunktionen des Merkmals „Nachname“ $f_1(N)$ und des Merkmals „Vorname“ $f_2(V)$ gilt damit:

$$(I) \qquad f_1(N) = \begin{cases} p(N), \text{ falls } p(N) \geq 0{,}5 \\ 0, \quad \text{sonst} \end{cases}$$

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

Die Bewertung der Übereinstimmung eines Datensatzpaares ergibt sich aus dem arithmetischen Mittel der Namens- und Vornamensähnlichkeit. Weiterhin erfolgt die Bewertung der Übereinstimmung und ggf. Klassifikation der Datensatzpaare in zwei Stufen:

In der ersten Stufe (I) erhalten Datensatzpaare nur dann eine Bewertung, wenn das arithmetische Mittel der Na-

mensähnlichkeiten mindestens 0,75 beträgt. In der zweiten Stufe (II) wird auf diese Nebenbedingung verzichtet. Unter Berücksichtigung der unterschiedlichen Schwellenwerte der beiden sukzessive durchlaufenden Stufen (I) und (II) lassen sich die Bewertungsfunktionen für ein Datensatzpaar (a,b) formal wie folgt darstellen:

(I)

$$\lambda(f_1(N), f_2(V)) = \begin{cases} 0{,}5\,(f_1(N)+f_2(V)), & \text{falls } 0{,}5\,(f_1(N)+f_2(V)) \geq 0{,}75 \text{ und } f_1(N), f_2(V) > 0 \\ 0, & \text{sonst} \end{cases}$$

(II)

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

Es wird das Datensatzpaar als identisch klassifiziert, das die höchste Bewertung der Übereinstimmung (>0) erzielt. Die Entscheidungsfunktion für beide Stufen lautet also:

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

mit λ_i (i=1,…,n) als den Bewertungen aller in den Vergleich einbezogenen Datensatzpaare.

Die Arbeitsweise von Unterstufe 215 soll an einem stark vereinfachten Beispiel demonstriert werden:

Der (noch nicht verknüpfte) Wohnungsinhabername in der GWZ lautet:

JOHANNES ZIMMERMANN

Im Melderegister finden sich noch folgende unverknüpfte Personendatensätze (b_1,b_2,b_3):

JOHAN SEMMERMANN, HANS ZIMMERER, HANNES ZIMMER

Das Phonetikmodul liefert uns folgende Ähnlichkeitsmessungen:

Paar	GWZ		Melderegister		P(V)	P(N)
	Vorname	Name	Rufname	Familienname		
(a,b_1)	**JOHAN**NES	ZI**MMERMANN**	**JOHAN**	SE**MMERMANN**	0,625	0,80
(a,b_2)	JO**HAN**NES	**ZIMMER**MANN	**HAN**S	**ZIMMER**ER	0,375	0,60
(a,b_3)	JO**HANNES**	**ZIMMER**MANN	**HANNES**	**ZIMMER**	0,75	0,60

Die Anwendung der o.g. Formeln führt zu folgendem Ergebnis:

Paar	$f_1(V)$	$f_2(N)$	Stufe I		Stufe II	
			$\lambda(f_1(N), f_2(V))$	$\delta(\lambda)$	$\lambda(f_1(N), f_2(V))$	$\delta(\lambda)$
(a,b_1)	0,625	0,80	0	nicht identisch	0,7125	Identisch
(a,b_2)	0	0,60	0	nicht identisch	0	nicht identisch
(a,b_3)	0,75	0,60	0	nicht identisch	0,675	nicht identisch

Somit wird der Name Johannes Zimmermann mit Johan Semmermann zusammengeführt.

Um die Leistungsfähigkeit des entwickelten Verfahrens bei einer empirischen Anwendung feststellen zu können, wurde es, noch ehe Daten aus dem Zensustest vorlagen, getestet. Dabei kam es zu folgenden Ergebnissen.

Testergebnisse

Für den Test standen Melderegisterdaten aus einigen Gemeinden der alten und neuen Bundesländer zur Verfügung. Statt der im Zensustest vorgesehenen Datensätze aus der Gebäude- und Wohnungszählung wurden Mikrozensusdatensätze aus den entsprechenden Gemeinden herangezogen. Vor Beginn des Tests stellte ein menschlicher Experte fest, wie viele der in den Daten des Mikrozensus vorhandenen Wohnungsinhaber sich auch in den Melderegisterdaten wiederfanden und auf welcher Stufe das zu testende Verfahren diese zusammenführen müßte. Anschließend wurde der maschinelle Namensabgleich gestartet und dessen Ergebnisse mit denjenigen des menschlichen Experten verglichen. Die nachfolgende Aufstellung gibt einen Überblick über den Verlauf des Tests:

Anzahl der Melderegistersätze: 2 962

Anzahl der Mikrozensussätze (Wohnungsinhaber): 1 772

Anzahl der in den Melderegistern vorhandenen Wohnungsinhaber (paarige Fälle): 1 683

Anzahl der nicht in den Melderegistern vorhandenen Wohnungsinhaber (Fehlbestände): 89

Zusammenführung/ Nichtzusammenführung	Anzahl der durch einen menschlichen Experten zusammengeführten Datensätze (paarige Fälle)	Davon		Trefferquote in % = Anteil der durch das Verfahren richtigerweise zusammengeführten paarigen Fälle an der Anzahl der paarigen Fälle insgesamt
		Anzahl der durch das Verfahren richtigerweise zusammengeführten paarigen Fälle	Anzahl der durch das Verfahren fälschlicherweise nicht zusammengeführten paarigen Fälle	
Insgesamt	1683	1671	12	99,3
Davon auf Hauptstufe 1	1564	1563	1	99,9
2	101	101	0	100,0
3	4	4	0	100,0
4	3	3	0	100,0
Auf weiterer im Test nicht angewendeter Stufe des Verfahrens	11	0	11	0,0

Die erzielten Ergebnisse zeigen, daß das Verfahren beinahe sämtliche durch einen menschlichen Experten herausgefundenen paarigen Fälle auch zusammenführte. Lediglich in 12 von 1683 paarigen Datensätzen war das nicht der Fall. In 11 dieser 12 Datensätze hätte die Zusammenführung allerdings auf Stufen des Namensabgleichs erfolgen müssen, die nicht Bestandteil des durchgeführten Tests waren. Deshalb können diese 11 Fälle nicht als Fehler des getesteten Verfahrens gewertet werden. Damit wurde nur ein einziger Datensatz fälschlicherweise nicht zusammengeführt.

Da bereits mit den im Test angewendeten Stufen des Namensabgleichs 99,3% der paarigen Datensätze zusammengeführt werden konnten, dürfte es bei Verwendung aller Stufen nach vorsichtiger Schätzung im Zensustest möglich sein, ca. 99,5% der Wohnungsinhaber maschinell mit den entsprechenden Melderegisterdaten zusammenzuführen. Für die verbleibende Restmasse von 0,5% ist eine manuelle Verknüpfung in einem Dialogverfahren vorgesehen.

Neben der korrekten Zusammenführung paariger Datensätze muß sich die Leistungsfähigkeit des Namensabgleichprogramms auch dadurch zeigen, daß Wohnungsinhaber, die nicht in den Melderegistern vorhanden sind (Fehlbestände), nicht fälschlicherweise mit irgendwelchen anderen Personen aus dem Melderegister zusam-

mengeführt werden. Bei den 89 Fehlbeständen des Testmaterials kam ein solcher Fehler nicht ein einziges Mal vor.

Zusammenfassend betrachtet, verlief der durchgeführte Test also recht vielversprechend. Allerdings wurde der Test mit kleinen Datenmengen durchgeführt, mit denen noch nicht alle Einzelheiten des Zusammenführungsverfahrens überprüft werden konnten. Mit dem sehr viel umfangreicheren Datenmaterial aus dem Zensustest müssen daher noch weitere Untersuchungen erfolgen, um das Verfahren im Hinblick auf seine mögliche Anwendung im Rahmen eines künftigen registergestützten Zensus noch hinreichend verfeinern zu können.

4. Fazit

Soll der Registerstatistik künftig eine größere Bedeutung zukommen als bislang, muß auch der Einsatz maschineller Verfahren zur Zusammenführung von Daten zwangsläufig deutlich ausgeweitet werden. Das im Aufsatz beschriebene, innerhalb des Zensustests entwikkelte Verfahren wurde auf die speziellen Erfordernisse der Zusammenführung von Melderegisterdaten mit denen der Gebäude- und Wohnungszählung zugeschnitten. Der Abgleich erfolgt im wesentlichen über die Namen natürlicher Personen und orientiert sich an den speziellen Randbedingungen der beiden Datenquellen. Das Verfahren kann daher nicht ohne weiteres für die Zusammenführung von Daten im Rahmen anderer Statistiken angewendet werden.

Die Grundidee des vorgestellten Abgleichverfahrens, der Zeichenkettenvergleich und das hierarchische Stufenmodell zur Bewertung, ist jedoch im Prinzip auch auf andere Zusammenführungsprobleme anwendbar. Mögliche weitere Anwendungsgebiete könnten Zusammenführungen von Datenbeständen anhand von Adressangaben oder Firmennamen sein. Für solche Anwendungen müßten allerdings in Analogie zum Zensustest jeweils eigene Verfahren mit speziellen Stufenfolgen konzipiert werden, die auf die Datenstruktur und Datenqualität der dann zusammenzuführenden Dateien auszurichten wären.

Dr. Michael Fürnrohr
Dipl.-Volksw. Birgit Rimmelspacher
Dipl.-Volksw. Tilman von Roncador

---

1) siehe hierzu Winkler, N.: Registernutzung im Rahmen einer Primärstatistik, Bayern in Zahlen, Heft 11/2000, S. 457ff.

2) Urteil des Bundesverfassungsgerichts vom 15. Dezember 1983 zum Volkszählungsgesetz 1983, Bundesanzeiger, Hrsg.: Bundesminister der Justiz, Nr. 241a, Jahrgang 35, 24. Dezember 1983.

3) Lenz, H. J., Neiling, M.: Data Integration by means of Object Identifikation in Information Systems. In Oroceedings of European Conference on Information Systems, Vienna, Austria, 2000.

4) Ausführliche Darstellungen des Zensustests finden sich u.a. bei Fürnrohr, M., Rimmelspacher, B.: Testuntersuchungen zur Vorbereitung eines registergestützten Zensus – Inhalt und Stand –, Bayern in Zahlen, Heft 1/2001, S. 13ff. sowie bei Lauer, T., Werner, J.: Der Zensustest 2001 – Prüfung neuer Methoden als Alternative für eine Volkszählung, Baden-Württemberg in Wort und Zahl, Heft 11/2001, S. 545 ff.

5) In der GWZ werden Wohnungsmerkmale, wie Anzahl der Räume, Wohnungsfläche oder Heizungsart von den Gebäude- und Wohnungseigentümern postalisch erfragt. Diese Befragung ist erforderlich, da es keine landesweiten Register mit Wohnungsangaben gibt.

6) Eine Beschreibung der Grundzüge des Verfahrens der Zusammenführung/Haushaltegenerierung findet sich bei: Fürnrohr, M., König, M.: Möglichkeiten einer Haushaltegenerierung im Rahmen der Zusammenführung von Einzeldaten aus Melderegistern mit primärstatistisch gewonnenen Wohnungsdaten, Bayern in Zahlen, Heft 4/1999, S. 161ff.

7) Vgl. hierzu www.uni-oldenburg.de/nausa/soundex.htm und www.javamentor/knowhow/algo-soundex.html am 7. 12. 2001.

8) Vgl. hierzu www.php-ressource.de/php/function.soundex.htm am 7. 12. 2001.

9) Unterstellt man, daß jeder Datensatz der GWZ im Melderegister gefunden wird, dann reduziert sich die Zahl der Vergleiche auf n(m-0,5(n-1)). Je nach der Größe von n und m beträgt die Reduktion der Vergleichsoperation bis zu knapp 50 Prozent.

Kleine Mitteilungen

Anbau von Ackerfrüchten in Bayern im Frühjahr 2002

Nach den Meldungen der rund 1300 amtlichen Ernteberichterstatter des Bayerischen Landesamts für Statistik und Datenverarbeitung vom April 2002 haben die landwirtschaftlichen Kulturen die winterliche Ruhephase überwiegend gut überstanden. Dennoch wiesen die Mehrzahl der Feldfrüchte und das Grünland Mitte April noch einen ungünstigeren Wachstumsstand als die Bestände im Vorjahr auf. Neubestellungen waren jedoch nur in geringem Umfang erforderlich.

Aus den hochgerechneten Angaben von knapp 1100 Betriebsberichterstattern ergibt sich gegenüber dem Vorjahr eine Ausdehnung der Ackerfläche um 0,9% auf 2105000 Hektar (ha). Dabei hat die Anbaufläche der Wintergetreidearten um 0,4% auf 897000 ha zugenommen. Der flächenmäßig bedeutende Winterweizen (455000 ha) erfuhr eine Ausdehnung um 0,2% und die Wintergerste (312000 ha) um 1,8%.

Die Anbaufläche der Sommergetreidearten wurde dagegen eingeschränkt, und zwar um 2,1% auf 210000 ha. Deutlich zurückgegangen ist auch der Anbau von Hafer