DESTATIS
wissen.nutzen.

Auszug aus Wirtschaft und Statistik

Herausgeber: Statistisches Bundesamt, Wiesbaden

Schriftleitung: Johann Hahlen
Präsident des Statistischen Bundesamtes
Verantwortlich für den Inhalt:
Brigitte Reimann,
65180 Wiesbaden

- Telefon: +49 (0) 6 11/75 20 86
- E-Mail: wirtschaft-und-statistik@destatis.de

Vertriebspartner: SFG Servicecenter Fachverlage
Part of the Elsevier Group
Postfach 43 43
72774 Reutlingen
Telefon: +49 (0) 70 71/93 53 50
Telefax: +49 (0) 70 71/93 53 35
E-Mail: destatis@s-f-g.com

Erscheinungsfolge: monatlich

Allgemeine Informationen über das Statistische Bundesamt und sein Datenangebot erhalten Sie:

- im Internet: www.destatis.de

oder bei unserem Informationsservice
65180 Wiesbaden

- Telefon: +49 (0) 6 11/75 24 05
- Telefax: +49 (0) 6 11/75 33 30
- E-Mail: info@destatis.de

Dr. Josef Schürle

Automatisierte Zusammenführung von Daten – Das Modell von Fellegi und Sunter

Josef Schürle wurde im Jahr 2004 für seine an der Universität Tübingen entstandene Dissertation „Record Linkage – Zusammenführung von Daten auf Basis des Modells von Fellegi und Sunter" ein Förderpreis des Statistischen Bundesamtes in der Kategorie „Dissertationen" des Gerhard-Fürst-Preises zuerkannt. Anknüpfend an die in den vergangenen beiden Ausgaben dieser Zeitschrift präsentierten Beiträge der Gerhard-Fürst-Preisträger von Gaudecker und Schürmann erläutert Josef Schürle im Folgenden die Inhalte seiner ausgezeichneten Arbeit.

1 Einleitung

Die automatisierte Zusammenführung von Daten aus unterschiedlichen Datenquellen gewinnt zunehmend an Bedeutung. Ziel ist es, Elemente zu identifizieren, welche in den verschiedenen Datensätzen gemeinsam enthalten sind. Dabei sind die potenziellen Anwendungsgebiete vielfältig. So kann man einerseits daran interessiert sein, verschiedene Datenbestände zu einem zusammenzufassen, um die weitere Verwaltung zu vereinfachen. Oder aber man möchte durch die Zusammenführung von Daten zusätzliche Informationen gewinnen. Beispielsweise können im Bereich der Medizin durch die Zusammenführung von alten mit neuen Daten Erkenntnisse über die Ursachen und die Entwicklung einzelner Krankheiten gewonnen werden. Eine primärstatistische Erhebung ist hier aus der Natur der Sache heraus vielfach nicht möglich. Im Bereich der deutschen amtlichen Statistik ist die Datenzusammenführung derzeit innerhalb der Mehrfachfallprüfung im Rahmen des Zensustests von Bedeutung.[1])

Für die automatisierte Zusammenführung von Daten existiert eine Vielzahl so genannter Ad-hoc-Ansätze. Ein einfaches Beispiel wäre, zwei Einheiten dann als identisch einzustufen, wenn sie bezüglich vier von fünf Vergleichsmerkmalen übereinstimmen. Aber auch ausgefeiltere Methoden sind im Einsatz. Derartige Methoden funktionieren im Einzelfall durchaus sehr gut, besitzen allerdings einige grundsätzliche Nachteile. Neben der Subjektivität der Ansätze ist vor allem ein Problem, dass diese wesentlich von den vorliegenden Daten abhängen. Eine allgemeine Beurteilung wird dadurch von vornherein ausgeschlossen. Insbesondere fällt auch eine Abschätzung der mit den Verfahren verbundenen Fehlerhäufigkeiten schwer.

Im Gegensatz dazu bietet das im Jahr 1969 von Ivan P. Fellegi und Alan B. Sunter vorgestellte Modell einen Ansatz, alle die genannten Nachteile zu eliminieren.[2]) Das auf wahrscheinlichkeitstheoretischen Überlegungen beruhende Modell wurde und wird insbesondere in den Vereinigten Staaten und in Kanada vielfältig eingesetzt.[3]) Im Weiteren wird das Modell in seiner ursprünglichen Form und eine Deutung im Sinne der klassischen Testtheorie skizziert. Anschließend werden zwei Ansätze zur Parameterschätzung dargestellt und die Ergebnisse einer Simulationsstudie diskutiert.

1) Siehe Lauer, T./Braun, R.: „Der Zensustest 2001 – Eine Zwischenbilanz aus ablauftechnischer und organisatorischer Sicht", Baden-Württemberg in Wort und Zahl, Heft 9/2002, S. 434 ff.

2) Siehe Fellegi, I. P./Sunter, A. B.: "A Theory for Record Linkage", Journal of the American Statistical Association, Band 64, 1969, S. 1183 ff.

3) Siehe Kilss, B./Alvey, W. (Hrsg.): "Record Linkage Techniques – 1985", Proceedings of the Workshop on Exact Matching Methodologies in Arlington, Virginia 1985 (*http://www.fcsm.gov/working-papers/RLT_1985.html*). Alvey, W./Jamerson, B. (Hrsg): "Record Linkage Techniques – 1997", Proceedings of an International Workshop and Exposition in Arlington, Virginia 1997 (*http://www.fcsm.gov/working-papers/RLT_1997.html*).

2 Das Modell von Fellegi und Sunter und seine Interpretation im Sinne der klassischen Testtheorie

Im Rahmen des Modells von Fellegi und Sunter wird die Menge aller aus zwei Datensätzen A und B zu bildenden Paare mit $A \times B$ bezeichnet und diese wiederum in die Mengen $M := \{(a,b) \in A \times B | a = b\}$ und $U := \{(a,b) \in A \times B | a \neq b\}$ zerlegt. Hierbei handelt es sich also um die Mengen aller aus identischen bzw. nicht-identischen Elementen gebildeten Paare. Ziel ist es, diese Mengen möglichst zuverlässig maschinell zu identifizieren. Hierfür werden drei mögliche Entscheidungen vorgegeben, nämlich die Zuordnung zu M (Entscheidung E_1), die Zuordnung zu U (Entscheidung E_3) und als neutrale Entscheidung die Nicht-Zuordenbarkeit (Entscheidung E_2). Nicht zugeordnete Elemente müssen nachträglich manuell ausgewertet werden. Jedes Paar aus $A \times B$ wird individuell betrachtet und die in beiden Datensätzen gemeinsam enthaltenen Merkmale werden verglichen. Als Ergebnis dieses Vergleichs werden jedem Paar für die drei möglichen Entscheidungen Wahrscheinlichkeiten zugeordnet und anschließend wird mittels eines Zufallsexperiments eine Entscheidung getroffen. Die Zuordnungsvorschrift, nach welcher die Wahrscheinlichkeiten vorgegeben werden, wird als (zufällige) Entscheidungsfunktion bezeichnet.

Das Optimalitätskriterium von Fellegi und Sunter orientiert sich an der Reduktion des nachträglichen Aufwandes, wobei die dabei resultierenden Fehler in vorgegebenen Grenzen gehalten werden sollen. Somit lautet der Optimierungsansatz für die zu wählende Entscheidungsfunktion:

$$\min P(E_2)$$

unter den Nebenbedingungen: $P(E_1|U) \leq \mu$ und $P(E_3|M) \leq \lambda$.

Auf Basis dieses Ansatzes ist es möglich, unter vergleichsweise schwachen Annahmen eine optimale Entscheidungsregel abzuleiten. Für die Festlegung von μ und λ sind noch so genannte Zulässigkeitsbedingungen zu beachten, die sicherstellen, dass die Entscheidungsfunktion wohl definiert ist.

Der Ansatz lässt sich in den Rahmen der klassischen Testtheorie nach Neyman und Pearson einbetten. Dazu werden die beiden einfachen Hypothesen

$$H_0 : (a,b) \in M \text{ und } H_1 : (a,b) \in U$$

aufgestellt. Für die Tests H_0 gegen H_1 bzw. H_1 gegen H_0 lassen sich gemäß dem Lemma von Neyman und Pearson[4]) jeweils beste Tests angeben. Sofern sich die kritischen Regionen und Randomisierungsbereiche der beiden Tests nicht überschneiden, führen diese bei unabhängiger Anwendung zu widerspruchsfreien Ergebnissen. Genau diesen Zweck erfüllen die bereits oben erwähnten Zulässigkeitsbedingungen. Durch geeignete Kombination der beiden Tests erhält man als Ergebnis die optimale Entscheidungsfunktion von Fellegi und Sunter.

Bei näherer Betrachtung stellt man fest, dass sich die Menge der zulässigen Signifikanzniveaus noch vergrößern lässt. Die besten Tests sind so konstruiert, dass die Randomisierungsbereiche jeweils einelementig sind. Addieren sich die Randomisierungswahrscheinlichkeiten zu einer Zahl kleiner oder gleich eins, so können auch noch μ- und λ-Werte zugelassen werden, für welche die Randomisierungsbereiche identisch sind. Sei t_λ der konstruierte beste Test für H_0 gegen H_1 und t_μ der Test für H_1 gegen H_0, so lässt sich zeigen[5]), dass die Festlegung

$$(P(E_1), P(E_2), P(E_3)) := (t_\mu, 1-t_\mu - t_\lambda, t_\lambda)$$

eine optimale Wahl im Sinne des von Fellegi und Sunter verfolgten Optimierungsansatzes ist, wobei die Menge der zulässigen μ- und λ-Werte im Vergleich zur ursprünglichen Darstellung vergrößert ist.

Der Ansatz lässt sich anschaulich deuten. Die Wahrscheinlichkeitstheorie wird dazu verwendet, den Informationsgehalt der verglichenen Merkmale zu messen. Reicht dieser aus, so führt dies zu einer der beiden Entscheidungen E_1 oder E_3. Ist der Informationsgehalt nicht ausreichend, so erhält man die Entscheidung E_2. Hierbei sind durchaus Parallelen zur menschlichen Logik zu erkennen. Je mehr Informationen zur Verfügung stehen und je besser diese sind, desto eher ist eine Entscheidung in die eine oder andere Richtung möglich. Da die Obergrenzen für die Fehlerwahrscheinlichkeiten in jedem Fall fixiert sind, führt ein zusätzlicher Informationsgehalt in den Merkmalen – unter sonst gleichen Voraussetzungen – zu einer geringeren Anzahl von E_2-Entscheidungen. Folglich spielen die vorhandenen Merkmale für die Qualität der Ergebnisse eine entscheidende Rolle, allerdings weniger für die Fehlerwahrscheinlichkeiten, sondern vielmehr für die Anzahl der resultierenden E_2-Entscheidungen.

3 Zwei Ansätze zur Schätzung der Modellparameter

Für die Konstruktion der Tests werden die Verteilungen der Vergleichsergebnisse unter H_0 und H_1 benötigt. Ein weit verbreiteter Ansatz besteht in der Maximum-Likelihood-(ML-) Schätzung unter Verwendung des so genannten EM-Algorithmus[6]). Hierbei handelt es sich um ein numerisches Verfahren, bei welchem ausgehend von einem vorzugebenden Startwert eine Parameter-Folge erzeugt wird, für welche die jeweiligen Funktionswerte der Likelihood-Funktion monoton gegen den Funktionswert eines stationären Punktes konvergieren. Die Voraussetzungen hierfür sind vergleichsweise schwach.[7])

4) Siehe Pruscha, H.: „Vorlesungen über Mathematische Statistik“, Stuttgart, Leipzig, Wiesbaden 2000, S. 222 ff.
5) Siehe Schürle, J.: „Record Linkage – Zusammenführung von Daten auf Basis des Modells von Fellegi und Sunter“, Frankfurt 2004, S. 34 f.
6) Siehe Dempster, A. P./Laird N. M./Rubin, D. B.: “Maximum Likelihood from Incomplete Data via the EM Algorithm”, Journal of the Royal Statistical Society, Series B, Band 39, 1977, S. 1 ff.
7) Siehe Wu, C. F. J.: “On the Convergence Properties of the EM Algorithm”, Annals of Statistics, Band 11, 1983, S. 95 ff.

Zunächst einmal muss spezifiziert werden, wie der Vergleich der vorhandenen Merkmale erfolgen soll. Bei den im Weiteren betrachteten Verfahren werden jeweils einfache „stimmt überein"/„stimmt nicht überein"-Vergleiche durchgeführt, was durchaus üblich ist. Dies kann zum einen dadurch erweitert werden, dass gewisse Verfahren, zum Beispiel phonetische Verfahren, vorgeschaltet werden, oder aber dadurch, dass die resultierenden Schätzwerte auf Basis einer so genannten Häufigkeitsadjustierung verallgemeinert werden.[8])

Im Wesentlichen unterscheiden sich die EM-basierten Schätzverfahren in der Spezifikation der Likelihood-Funktion und folglich in den zugrunde liegenden Annahmen. Treffen die Annahmen nicht zu, so sind die resultierenden Schätzwerte zwar ML-Schätzer im Sinne der verwendeten Funktion, es können aber bedeutende Verzerrungen im Vergleich zu realen Gegebenheiten auftreten. Hiervon wird die Qualität der Ergebnisse natürlich erheblich beeinflusst.

Eine weit verbreitete Spezifikation beruht auf der Annahme der bedingten Unabhängigkeit.[9]) Dabei wird unterstellt, dass die Vergleichsergebnisse für die einzelnen betrachteten Merkmale unter H_0 bzw. H_1 unabhängig voneinander sind. Dies kann zum Teil bezweifelt werden. So führt unter H_1 beispielsweise die Übereinstimmung bezüglich des Vornamens zu einer hohen Wahrscheinlichkeit der Übereinstimmung bezüglich des Geschlechts. Betrachtet man die Merkmale als unabhängig, so wird jede Übereinstimmung für sich gewertet und die Bedeutung somit überschätzt.

Ein alternativer Ansatz besteht darin, sämtliche Abhängigkeiten explizit zu modellieren.[10]) Dies führt dazu, dass diesbezüglich keine fehlerhaften Annahmen in die Modellierung eingehen, bringt aber einen anderen wesentlichen Nachteil mit sich. Aufgrund der starren Spezifikation der Likelihood-Funktion durchläuft der EM-Algorithmus nur eine Iteration bis zum Ergebnis. Es wird auf Basis der durch den Startwert implizierten Aufteilung in die Mengen M und U ein ML-Schätzer bestimmt. Folglich setzt diese Methode voraus, dass der verwendete Startwert eine möglichst zuverlässige Aufteilung liefert. Zur Bestimmung dessen kann ein vereinfachender Ansatz gewählt werden.[11]) Dazu muss man sich vor Augen führen, dass die wesentliche Schwierigkeit darin liegt, die Wahrscheinlichkeiten der Übereinstimmung der Merkmale unter H_1 zu bestimmen. Diese hängen von den betrachteten Merkmalen und von der vorliegenden Stichprobe ab. So liegt beispielsweise die Übereinstimmung bezüglich des Geschlechts bei unterschiedlichen Personen in der Nähe von 50%, im Falle des Vornamens liegt die Wahrscheinlichkeit deutlich niedriger. Wählt man Stichproben, die im Extremfall nur aus Männern bestehen, so liegt die Wahrscheinlichkeit innerhalb der Stichproben bei 100%. Im Gegensatz dazu stimmen die Merkmale bei ein und derselben Person mit einer Wahrscheinlichkeit nahe 100% überein. Natürlich hängt dies auch von dem Merkmal und den verwendeten Stichproben ab – ein Vorname wird seltener geändert als die Adresse und ein größerer zeitlicher Abstand zwischen den Erhebungszeitpunkten der einzelnen Stichproben führt zu größeren Veränderungen – die grundsätzlichen Zusammenhänge sind aber leichter abzuschätzen. Wird beispielsweise das Merkmal Name betrachtet, so kann der Anteil der Namensänderungen in dem betrachteten Zeitraum als Grundlage genommen werden und – nach einer Korrektur für Erhebungs- bzw. Angabefehler – als Startwert vorgegeben werden. Schätzt man noch den Anteil der Menge M an der Menge $A \times B$ – was im Sinne einer Wirtschaftlichkeitsbetrachtung in jedem Fall erfolgen sollte – so lassen sich die Wahrscheinlichkeiten unter H_1 auf Basis der Informationen in den Daten und unter Verwendung der Unabhängigkeitsannahme schätzen. Dies ist insoweit gerechtfertigt, als dass für den konstruierten EM-basierten Ansatz nicht der Startwert selbst, sondern die daraus resultierende Aufteilung in M und U wesentlich ist.[12])

4 Ergebnisse einer Simulationsstudie

Zur praktischen Erprobung der Schätzverfahren im Zusammenhang mit dem Modell von Fellegi und Sunter wurde eine ausführliche Simulationsstudie durchgeführt. Die Grundlage bildeten Telefondaten von Berlin, wobei nach einem Bereinigungsprozess 895 192 Personendatensätze zur Verfügung standen. Diese wurden eindeutig indiziert und dieser Index wurde in jedem Schritt mitgeführt, was eine vollständige Analyse der realisierten Ergebnisse ermöglichte. In Tabelle 1 sind die verwendeten Merkmale und die gewählten Merkmalskombinationen abgebildet.

Tabelle 1: Verwendete Merkmale und gewählte Merkmalskombinationen

Merkmale	Merkmalskombinationen
Name	V1 = Name
Vorname	V2 = V1 + Vorname
Straße	V3 = V2 + Straße
Hausnummer	V4 = V3 + Hausnummer
Postleitzahl	V5 = V4 + Postleitzahl
Ortsteil	V6 = V5 + Ortsteil
Anrede	V7 = V6 + Anrede

Es wurde unter verschiedenen Szenarien simuliert, wobei jeweils 250 Stichprobenpaare im Umfang von je 1 000 Einheiten je Stichprobe nach einem vorgegebenen Schema zufällig ausgewählt wurden. Zusätzlich wurden zufällige Fehler erzeugt. Analysiert wurden die Auswirkungen der Qualität und Quantität der Merkmale, des gewählten Startwertes, der Signifikanzniveaus und der Stichproben, einschließlich der vorhandenen Fehler, im Hinblick auf die rea-

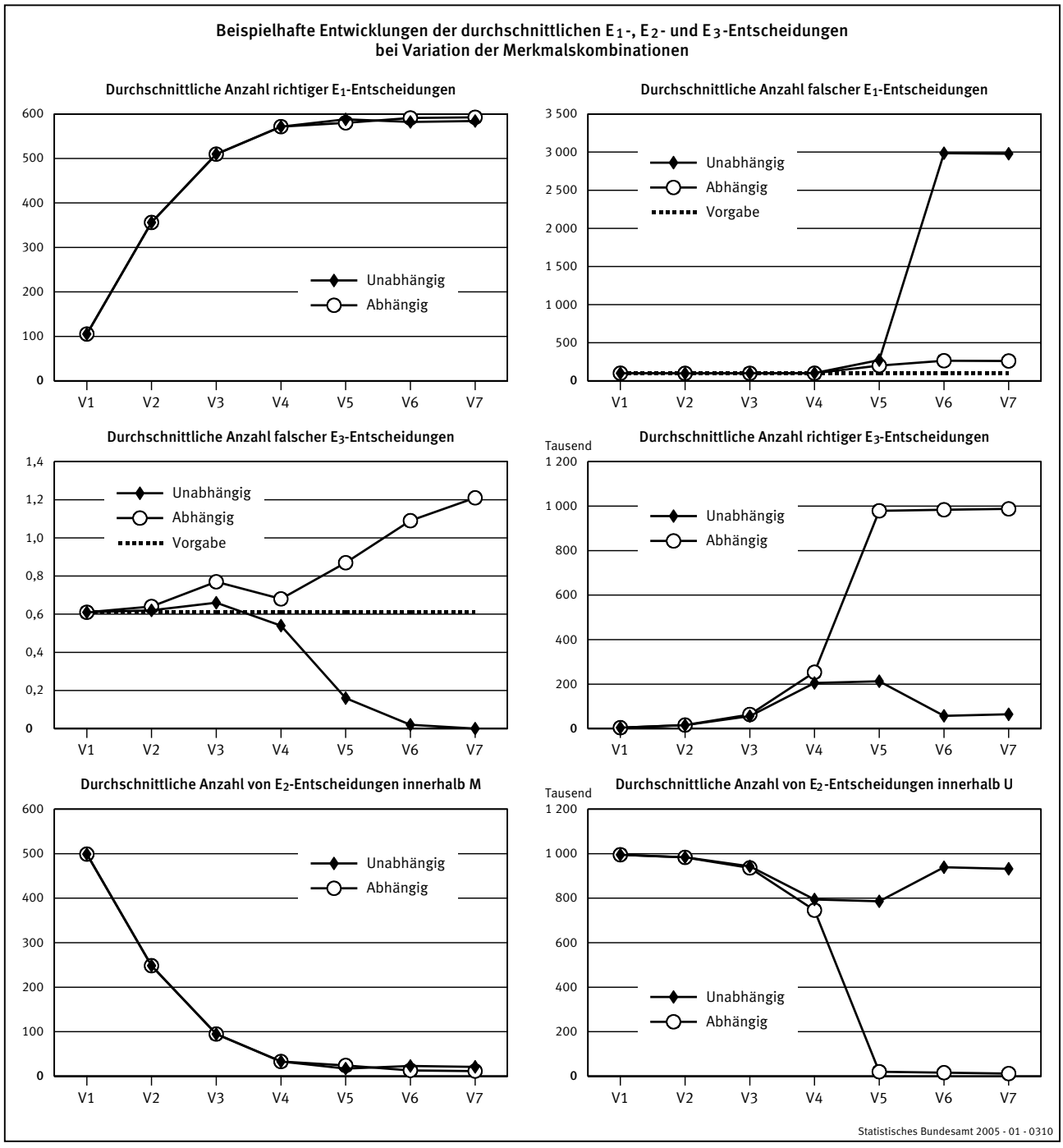
8) Siehe Fußnote 5, S. 102 ff.; Winkler, W. E.: "Frequency-Based Matching in Fellegi-Sunter Model of Record Linkage", Technical Report RR 00/06, U.S. Bureau of the Census, Statistical Research Division (*http://www.census.gov/srd/www/byyear.html*); Yancey, W. E.: "Frequency-Dependent Probability Measures for Record Linkage", Technical Report RR 00/07, U.S. Bureau of the Census, Statistical Research Division (*http://www.census.gov/srd/www/byyear.html*).

9) Siehe Winkler, W. E.: "Advanced Methods for Record Linkage", Technical Report RR 94/05, U.S. Bureau of the Census, Statistical Research Division (*http://www.census.gov/srd/www/byyear.html*).

10) Siehe Schürle, J.: "A Method for Consideration of Conditional Dependencies in the Fellegi and Sunter Model of Record Linkage", zur Publikation angenommen in Statistical Papers, 2003.

11) Siehe Fußnote 5, S. 99 ff.

12) Siehe Fußnote 5, S. 91 ff.



lisierten Ergebnisse. Einige wesentliche Erkenntnisse werden im Folgenden dargestellt.

Zunächst einmal zeigt sich, dass sich der Informationsgehalt einzelner Merkmale – wie zu erwarten – in den realisierten Ergebnissen auswirkt. So liefert der Name mehr Informationen als beispielsweise die Postleitzahl und führt somit auch zu besseren Resultaten. Wie bereits diskutiert, spiegeln sich die besseren Ergebnisse aber nicht in den realisierten Fehlerzahlen wider – diese sind ja über die Konstruktion des Modells fixiert –, sondern in einer deutlichen Reduktion der Anzahl der E_2-Entscheidungen. Dies zeigt wieder den Charakter der Informationsmessung, den dieses Modell besitzt. Ebenso zeigt sich in der Tendenz, dass der E_2-Bereich umso kleiner wird, je mehr Merkmale herangezogen werden. Allerdings gilt diese Aussage nicht generell, wie die beispielhaft im Schaubild dargestellten Ergebnisse verdeutlichen. Darin sind für ein Szenario die durchschnittlichen Anzahlen der richtigen und falschen E_1- und E_3-Entscheidungen, sowie die durchschnittlichen Anzahlen der realisierten E_2-Entscheidungen in Abhängigkeit von der verwendeten Merkmalskombination dargestellt. Gezeigt werden Ergebnisse unter Annahme von bedingter Unabhängigkeit („unabhängig“) und bei expliziter Modellierung der Abhängigkeiten („abhängig“).

Tabelle 2: Bei Paaren aus der Menge U durchschnittlich beobachteter Anteil der Übereinstimmung bezüglich einer Variablen bei Übereinstimmung bzw. Nicht-Übereinstimmung bezüglich einer anderen Variablen

Gegenstand der Nachweisung	Merkmale	Anteil der Übereinstimmung bezüglich der Variablen				
		Vorname	Straße	Postleitzahl	Ortsteil	Anrede
Übereinstimmung bezüglich der Variablen	Vorname	1	0,0005	0,0088	0,0221	**0,7419**
	Straße	0,0039	1	**0,4160**	**0,4715**	0,4193
	Postleitzahl	0,0036	**0,0201**	1	**0,3571**	0,4190
	Ortsteil	0,0035	**0,0087**	**0,1364**	1	0,4201
	Anrede	**0,0062**	0,0004	0,0085	0,0222	1
Nicht-Übereinstimmung bezüglich der Variablen	Vorname	0	0,0004	0,0085	0,0222	**0,4186**
	Straße	0,0035	0	**0,0083**	**0,0220**	0,4197
	Postleitzahl	0,0035	**0,0002**	0	**0,0193**	0,4197
	Ortsteil	0,0035	**0,0002**	**0,0056**	0	0,4197
	Anrede	**0,0016**	0,0004	0,0085	0,0222	0

Die Ursache für die schlechten Ergebnisse ab der Merkmalskombination V5 unter der Unabhängigkeitsannahme sind bedingte Abhängigkeiten, die tatsächlich in den Daten vorhanden sind. In Tabelle 2 sind relative Anteile der Übereinstimmung innerhalb der Daten bezüglich einiger Variablen unter der Voraussetzung, dass bezüglich einer anderen Variablen Übereinstimmung besteht, beispielhaft dargestellt. Hierbei wurde eine Beschränkung auf die Menge U vorgenommen, da es ja nicht auf Abhängigkeiten an sich, sondern auf bedingte Abhängigkeiten ankommt. Es zeigt sich, dass die Variablen Vorname und Anrede[13]) sowie Straße, Postleitzahl und Ortsteil zum Teil starke Abhängigkeiten aufweisen. Dies ist sicherlich nicht verwunderlich und wird durch die Daten nochmals belegt. Als Ergebnis führt die in diesem Fall falsche Unabhängigkeitsannahme zu starken Verzerrungen und damit zu zum Teil unbrauchbaren Resultaten. Bei expliziter Modellierung der Abhängigkeiten fallen diese deutlich weniger ins Gewicht. Die vorhandenen Informationen können dazu genutzt werden, die Anzahl der E_2-Entscheidungen nochmals zu reduzieren und somit das Resultat weiter zu verbessern.

Wie angemerkt, besteht die Problematik bei der Modellierung aller Abhängigkeiten in der Bedeutung des Startwerts. Aus diesem Grund wurde die Sensitivität der Qualität der Ergebnisse bezüglich Variationen des Startwerts untersucht. Es zeigt sich, dass die grundsätzliche Entwicklung der Ergebnisse bei unterschiedlichen Startwerten erhalten bleibt. Zusätzliche Informationen in den Daten führen zu einer Verbesserung der Ergebnisse und zu einer monotonen Reduktion des E_2-Bereichs. Allerdings verändern sich die absoluten Zahlen und folglich auch die realisierten Fehlerhäufigkeiten, wenn auch zum Teil nicht wesentlich. Die Untersuchungen zeigen aber auch, dass durch eine konservative Wahl des Startwerts – das heißt der Informationsgehalt der Merkmale wird bei der Vorgabe eher unter- denn überschätzt – eine Unterschreitung der vorgegebenen Fehlerhäufigkeiten erfolgt. Somit bleibt das Ergebnis weiterhin unter Kontrolle.

Eine weitere Erkenntnis lautet, dass eine größere Anzahl von Fehlern in den Daten zu einer Erhöhung der Anzahl der E_2-Entscheidungen führt. Durch die Fehler wird der Informationsgehalt der Merkmale tendenziell reduziert, was sich in dem beobachtbaren Phänomen äußert. Auch hierbei handelt es sich wiederum um ein zu erwartendes Ergebnis.

5 Fazit

Gegenüber herkömmlichen Ad-hoc-Verfahren besitzt der Ansatz von Fellegi und Sunter wesentliche Vorteile. Zunächst einmal ist hier die große Flexibilität der Methode zu nennen. Sie stellt nur geringe Anforderungen an die vorhandenen Daten und setzt keine spezielle Datensituation voraus. Eventuell in den Daten enthaltene Fehler werden implizit berücksichtigt. Die gewählte Entscheidungsfunktion hängt ausgehend von dem gewählten Kriterium nicht von subjektiven Gegebenheiten ab und ist bezüglich dieses Kriteriums optimal. Des Weiteren ermöglicht der Ansatz eine Fehlerkontrolle, da die tolerierbaren Fehlergrenzen unmittelbar in das Modell eingehen. Dem stehen allerdings höhere Anforderungen für den Anwender bzw. Entwickler gegenüber. Eine wesentliche Schwierigkeit besteht darin, die benötigten Verteilungen zu schätzen.

Insgesamt zeigen die betrachteten Verfahren zur Parameterschätzung unter idealen Bedingungen sehr gute Ergebnisse. Geht man von bedingter Unabhängigkeit aus, so führen vorhandene Abhängigkeiten hingegen zu Verzerrungen und somit zu zum Teil unbrauchbaren Resultaten. Modelliert man alle Abhängigkeiten explizit, so ist ein starker Einfluss des Startwerts die Folge. Allerdings zeigt sich, dass – unter Verwendung einer speziellen Prozedur zur Bestimmung des Startwerts – Informationen in den Daten in der richtigen Art und Weise verarbeitet und in bessere Ergebnisse umgesetzt werden. Geht man bei der Bestimmung des Startwerts nach diesem speziellen Verfahren und konservativ vor, so ist auch eine Einhaltung der Fehlerwahrscheinlichkeiten gewährleistet, das heißt die Ergebnisse bleiben weiterhin unter Kontrolle.

Die Untersuchungen verdeutlichen, dass das Modell von Fellegi und Sunter sowohl bezüglich seiner theoretischen Eigenschaften als auch bezüglich seiner praktischen Resultate sehr gut zur automatisierten Zusammenführung von Daten geeignet ist. Die größte Hürde für eine flächendeckende Anwendung besteht sicherlich darin, eine standardisierte Software zu entwickeln, die den Anforderungen eines Großteils der potenziellen Nutzer entspricht. Mit diesem Werkzeug wäre dann der wesentliche Nachteil – der große Aufwand – entscheidend reduziert.

13) Als mögliche Anrede kommen hierbei Herr und Frau in Betracht. Insofern kann das Merkmal mit dem Geschlecht gleichgesetzt werden.